2017

Next Generation Sequencing-Based Genotyping of Human Blood Groups: FY, JK and ABO Genes

Altayar, Malik Abdullah

Next Generation Sequencing-Based Genotyping of Human Blood Groups: *FY*, *JK* and *ABO* Genes

by

Malik Abdullah Altayar

A thesis submitted to Plymouth University
in partial fulfilment for the degree of

DOCTOR OF PHILOSOPHY

School of Biomedical and Healthcare Sciences

September 2016

Acknowledgements

First of all, I thank God for all his blessings that enable me to complete this PhD project. My sincere gratitude is to my parents and all my family for their great support. I am sincerely grateful for the support of my wife and daughters.

I would like to express my deepest gratitude to Tabuk University for funding my studies and providing great support in my scholarship in the United Kingdom.

I would like to express my special appreciation to my supervisors, Professor Neil Avent and Dr. Tracey Madgett for their endless support and guidance. It has been a great pleasure and honour to be part of their research team. I would like to thank Dr. Michele Kiernan, Dr. Kris Jeremy and Dr. Paul Waines for their valuable help. Finally, I would like to thank all my friends and anyone who have been supporting me throughout my scholarship.

AUTHOR'S DECLARATION

At no time during the registration for the degree of *Doctor of Philosophy* has the author been registered for any other University award without prior agreement of the Graduate Sub-Committee.

Name: Malik Abdullah Altayar

Signature:

Date:

Word count of the thesis's main body: 64,612 words

Award

Margaret Kenwright Young Scientist Award 2016 from the British Blood Transfusion Society (BBTS) for the following work:

Altayar, M.A., Madgett, T.E., Kiernan, M., Halawani, A.J., Avent, N.D. (2016). Next generation sequencing of *JK* (*SLC14A1*) gene reveals higher frequency of variant alleles, novel allele-defining SNPs (allele reference fingerprints) and reassignment of a purported *JKnull* allele. *Transfusion Medicine,* **26,** suppl. 2**,** 3-24.

Publications

N. D. Avent, T. E. Madgett, A. J. Halawani, **M. A. Altayar,** M. Kiernan, A. J. Reynolds, & X. Li. (2015). Next-generation sequencing: academic overkill or high-resolution routine blood group genotyping? *ISBT Science Series*, **10,** suppl. 1, 250-256.

Conference participation

M. Altayar, A. Halawani, M. Kiernan, A. Reynolds, N. Kaushik, T. Madgett & N. Avent (2013) Next Generation Sequencing of ABO, Duffy and Kidd Blood Group Genotyping. *Transfusion Medicine,* **23,** suppl. 2; 30-71.

A. J. Halawani, **Altayar, M. A.,** M. Kiernan, N. Kaushik, A. Reynolds, T. Madgett & N. Avent (2013) Comprehensive Genotyping for Kell and Rh Blood Group Systems by Next-generation DNA Sequencing. *Transfusion Medicine* **23,** suppl. 2; 30-71.

Altayar, M. A., Halawani, A. J., Kiernan, M., Reynolds, A. J., Kaushik, N., Madgett, T. E. & Avent, N. D. (2014). Extensive Genotyping of Blood Groups Duffy, Kidd and ABO by Next-generation Sequencing. *Vox Sang,* **107,** suppl. 1**,** 57-248.

Halawani, A. J., **Altayar, M. A.,** Kiernan, M., Reynolds, A. J., Kaushik, N., Madgett, T. E. & Avent, N. D. (2014). Can Next-generation DNA Sequencing Solve the RH Complexity for Genotyping? *Vox Sang,* **107,** suppl. 1**,** 57-248.

Altayar, M. A., Halawani, A. J., Kiernan, M., Madgett, T. E. & Avent, N. D. (2015). Complete Gene Sequencing of ABO Blood Group by Next-generation Sequencing. *Vox Sang,* **109,** suppl. 1**,** 1-379.

Halawani, A. J., **Altayar, M. A.,** Kiernan, M., Li, X., Madgett, T. E. & Avent, N. D. (2015). High Resolution Genotyping of the Rh Blood Group System by Next-generation Sequencing *Vox Sang,* **109,** suppl. 1**,** 1-379.

Abstracts are shown in Appendix A

Abstract

Next Generation Sequencing-Based Genotyping of Human Blood Groups: *FY, JK* and *ABO* Genes.

Malik Abdullah Altayar

Serological discrepancies in matching blood group antigens between donors and patients for blood transfusion may lead to alloimmunisation, especially in multiply transfused patients. Blood group genotyping (BGG) has contributed in reducing this issue. ABO, Fy and Jk antigens are among those to be causative for alloimmunisation through transfusion or pregnancy. The number of alleles of these clinically significant blood groups is ever increasing. Currently, all commercially available high-throughput BGG platforms are only based on pre-defined polymorphisms. Consequently, novel or rare alleles that might have clinical significance are not identified. Next generation sequencing (NGS) circumvents this issue by providing high-throughput comprehensive genotyping of blood group genes in discovery mode to find all existing and novel mutations. Accordingly, a large number of individuals can be genotyped in a single run. Here, we describe an NGS-based method coupled with long-range polymerase chain reaction (LR-PCR) for high-throughput, rapid and extensive genotyping of *FY*, *JK* and *ABO* blood group genes. The Ion Torrent Personal Genome Machine (PGM^{TM}) was used for sequencing the entire *FY*, *JK* and *ABO* blood group genes including flanking regions. Accordingly, high resolution genotyping was obtained. 53 genomic DNA samples were sequenced and genotyped for *FY*, 67 for *JK* and 47 for *ABO*. Sequencing data were aligned to the gene reference sequence derived from the human genome (hg19) to analyse variants. Analysis was accomplished by software packages, such as Ion Torrent Suite^{TM} plugins. Sanger sequencing of cDNA and cDNA clones was used to confirm findings in the *JK* gene. The sequencing data had a coverage depth of more than 5000x for *FY*, 700x for *JK* and 600x for *ABO*. NGS data matched with the serological phenotypes of *FY* alleles *FY*A, FY*B* and *FY*02 Null* main polymorphisms, such as *FY*A*/*FY*B* (125G>A) in exon 2 and (-67 T>C) in the promotor region. *JK* variant analysis revealed that the *JK*01W.01* allele (130G>A) is common (10/67 samples) with normal antigenicity. The previously described silencing polymorphism (810G>A), leading to a purported *JK*B* null allele, restores a splice site and does not correlate with loss of Jk^{b} antigenicity (10/67 samples). *JK* intron analysis revealed several new *JK* alleles described in this thesis. All 7 exons, introns and the flanking regions of the *ABO* gene were covered by only four amplicons. Several rare *O* alleles were found, such as *O73* and *O75*, while one suggested novel *O* allele was characterised by a missense SNP 482G>A (Arg161His) in exon 7. The *ABO* reference sequence from hg19 appeared to resemble (*O01* and *O02*) alleles. The intronic SNPs might be used to distinguish between alleles more accurately as a correlation of the intronic SNPs with the alleles was noted for the homozygous *O* alleles. It is predicted that NGS-based genotyping will replace not only microarray-based genotyping but also serology in the blood group typing of individuals, with great advancements in technology and molecular knowledge being expected in the near future.

Table of Contents

Chapter 1: Introduction	1
1.1 Introduction to blood groups	1
1.2 Blood group terminology	2
1.3 The molecular basis of blood group polymorphisms	6
1.4 The immunogenicity of blood groups	10
1.4.1 Haemolytic transfusion reaction (HTR)	10
1.4.2 Haemolytic disease of fetus and newborn (HDFN)	11
1.5 ABO blood group system	12
1.5.1 Discovery	13
1.5.2 Biosynthesis of ABO blood group antigens	15
1.5.3 Inheritance and molecular genetics	18
1.5.4 *ABO* polymorphisms (*ABO* alleles)	21
1.5.5 *ABO* antibodies (clinical significance)	24
1.6 Duffy blood group system (FY)	25
1.6.1 Duffy glycoprotein and *FY* gene.	25
1.6.2 Molecular basis of FY blood group system	27
1.6.3 FY antibodies (clinical significance)	31
1.7 Kidd blood group system (JK)	31
1.7.1 JK glycoprotein and the *JK* gene	32
1.7.2 Molecular basis of JK blood group system	34
1.7.3 JK antibodies (clinical significance)	35
1.8 Blood group genotyping	36
1.8.1 Serology and genotyping	36
1.8.2 Applications of BGG	38
1.8.3 High throughput	39
1.8.4 Genotyping technology and methodology	40
1.8.5 Next-generation sequencing	45
1.8.6 Ion Torrent Personal Genome MachineTM (Ion PGM^{TM})	49
1.8.7 NGS and BGG	54
1.9 Thesis aims	56
CHAPTER 2: Materials and Methods	58
2.1 Genomic DNA (gDNA) and RNA from whole blood samples	58
2.2 Sequencing of blood group genes by next-generation sequencing	58
2.2.1 gDNA extraction and purification	61
2.2.2 RNA extraction and purification	62
2.2.2.1 RNA quantification	64
2.2.3 DNA quantitation	65
2.2.4 Amplicon library preparation for next-generation sequencing	66
2.2.4.1 Primers	66
2.2.4.2 LR-PCR amplicon amplification	70
2.2.4.3 Agarose gel electrophoresis	72
2.2.4.4 Purification	74

2.2.4.5 Sample aliquots 77
2.2.4.6 Amplicon library fragmentation (purified fragmented library) 78
2.2.4.7 The Agilent® 2100 Bioanalyzer 80
2.2.4.8 Optimisations of the fragmentation reaction time 83
2.2.4.9 Ligation of barcoded adapters (purified-ligated library) 87
2.2.4.10 Size selection of the library 90
2.2.4.11. Library quantitation using 2100 Bioanalyzer® instrument for Ion Template preparation. 96
2.2.5 Template preparation 97
2.2.6 Sequencing 100
2.2.7 Data analysis and bioinformatics 101
2.3 *JK*: cDNA analysis of G810A (exon 8 near splice region) and A588G (exon 7 with *JK*01W*). 102
2.3.1 (*JK*) cDNA analysis of G810A 102
2.3.2 (*JK*) cDNA analysis of A588G 104
2.3.3 Primers 105
2.3.4 cDNA synthesis and specific amplification 107
2.3.5 Amplification of target cDNA 108
2.3.6 cDNA Sequencing (Sanger) of amplicon covering G810A 110
2.3.7 Molecular cloning 112
2.3.7.1 Isolation and purification of plasmid DNA 113
2.3.7.2 Restriction enzyme analysis to screen plasmid DNA for insert 114
2.4 Sanger sequencing for *ABO* gene 115
2.5 Sanger sequence analysis 116
Chapter 3: Genotyping of the FY Blood Group by Next Generation Sequencing 117
3.1 Introduction 117
3.2 Aim of the study 119
3.3 Results 120
3.3.1 Long-range PCR of the *FY* gene 120
3.3.2 Next Generation Sequencing of the *FY* gene 124
3.3.2.1 *FY* amplicon library fragmentation (purified fragmented library) 124
3.3.2.2 Ligation of barcoded adapters (*FY* purified-ligated library) 125
3.3.2.3 Size selection 126
3.3.3 Next Generation Sequencing data quality control 132
3.3.3.1 Sequencing Data summary report 132
3.3.3.2 NGS data quality control 136
3.3.3.2.1 Per base sequence quality 136
3.3.3.2.2 Per sequence quality scores 137
3.3.3.3 NGS sequence visualisation 141
3.3.3.4. Variant analysis and Genotyping 143
3.3.4 *FY* genotyping results from Next Generation Sequencing 144
3.3.4.1 Mutations in exons and the promoter region 145
3.3.4.2 Mutations in the intron and downstream region. 146

3.4 Discussion 151
3.4.1 Next Generation Sequencing of the *FY* library 151
3.4.2 Quality control of Next Generation Sequencing data 151
3.4.3 Validity of *FY* Next Generation Sequencing genotyping 154
3.4.4 *FY* Next Generation Sequencing genotyping 154
3.4.4.1 Mutations in exons and the promoter region 156
3.4.4.2 Mutations in introns 157
3.4.5 The advantage of Next Generation Sequencing genotyping of *FY* 159
Chapter 4: Genotyping of the JK Blood Group by Next Generation Sequencing 162
4.1 Introduction 162
4.2 Aim of the study 163
4.3 Results 164
4.3.1 LR-PCR of the *JK* gene 164
4.3.2 NGS the *JK* gene 168
4.3.2.1 *JK* amplicon library fragmentation (purified-fragmented library) 168
4.3.2.2 Ligation of barcoded adapters (*JK* purified-ligated library) 170
4.3.2.3 Size selection 170
4.3.3 NGS data quality control 173
4.3.3.1 Sequencing data summary report 173
4.3.3.2 NGS data quality control 175
4.3.3.2.1 Per base sequence quality 175
4.3.3.2.2 Per sequence quality scores 175
4.3.3.3 NGS sequence visualisation 179
4.3.3.4 Variant analysis and genotyping 181
4.3.4 *JK* genotyping by NGS 182
4.3.4.1 NGS analysis and Genotyping of the *JK* gene 182
4.3.4.2 Assignment of *JK* allele-specific polymorphism patterns 185
4.3.5 cDNA analysis of *JK* 810G>A and 588A>G SNPs 192
4.3.5.1 SNP 810G>A 192
4.3.5.2 SNP 588A>G 197
4.4 Discussion 201
4.4.1 NGS of the *JK* library 201
4.4.2. Quality control of NGS data 201
4.4.3 Validity of *JK* NGS genotyping. 202
4.4.4 *JK* NGS-based genotyping 203
4.4.4.1 Genotyping of *JK* SNPs in exons 205
4.4.4.2 *JK* polymorphism patterns and assignment of *JK*A*, *JK*B* and *JK*01W* allele reference sequences (‘fingerprints’) 207
Chapter 5: Genotyping of the ABO Blood Group by Next Generation Sequencing 210
5.1 Introduction 210
5.2 Aim of the study 211
5.3 Results 212

5.3.1 LR-PCR of the *ABO* gene 212
5.3.2 NGS of the *ABO* gene 216
5.3.2.1 *ABO* purified fragmented library 216
5.3.2.2 Ligation of barcoded adapters (*ABO* purified-ligated library) 218
5.3.2.3 Size selection 218
5.3.3 NGS data quality control 221
5.3.3.1 Sequencing data summary report 221
5.3.3.2 NGS data quality control 223
5.3.3.2.1 Per base sequence quality 223
5.3.3.2.2 Per sequence quality score 225
5.3.4 NGS sequence visualisation 227
5.3.4.1 Observations on the *ABO* reference sequence 228
5.3.4.2 Observations on the *ABO* amplicons 232
5.3.5 Variant analysis and genotyping 234
5.3.6 *ABO* genotyping results from NGS 236
5.3.6.1 *ABO* mutations in exons 236
5.3.6.2 *ABO* mutations in introns 244
5.3.7 Validity of NGS data 255
5.4 Discussion 257
5.4.1 *ABO* NGS genotyping with LR-PCR 257
5.4.2 Quality control of NGS data 259
5.4.3 Validity of *ABO* NGS 260
5.4.4 NGS *ABO* library 260
5.4.5 NGS data analysis (*ABO* reference sequence) 261
5.4.6 Genotyping of the *ABO* gene 263
5.4.6.1 Polymorphisms in exons 263
5.4.6.2 Polymorphisms in introns and upstream regions 265
Chapter 6: General Discussion and Conclusion 269
6.1 NGS with LR-PCR 270
6.1.1 NGS data quality 272
6.2 Polymorphisms in Exons 272
6.3 The Intronic Polymorphisms 275
6.4 *Cis* or *Trans?* (Assignment of haplotype) 277
6.5 Data analysis 278
6.6 NGS costs 279
6.7 Future advancements in NGS technology 280
6.8 Future work 282
6.8.1 Allele classification system 283
6.9 Conclusion 285
References 286
Appendices 309

List of Figures

Figure 1.1 A brief diagram of A, B and H antigen (also O group) oligosaccharidesstructure biosynthesis.	17
Figure 1.2 Diagramatic illustration of the products (glycotransferases), of four *ABO* alleles, located in Golgi membrane.	19
Figure 1.3 The ABO gene (seven exons and six introns).	20
Figure 1.4. The structure of the Duffy protein (7-transmembrane domain).	30
Figure 1.5 The Kidd glycoprotein, showing 10 membrane spanning domains.	33
Figure 1.6 Ion PGMTM sequencing technology and chemistry.	52
Figure 1.7 Summary of the Ion PGMTM workflow.	53
Figure 2.1. Overview of the process for sequencing blood group genes by Next-Generation Sequencing.	60
Figure 2.2. Separation of blood components in a tube.	65
Figure 2.3. GeneRulerTM 1Kb Plus DNA ladder (Thermo Fisher Scientific) used as a marker of DNA size	73
Figure 2.4. The principle of the purification mechanism using magnetic beads.	76
Figure 2.5. Example Bioanalyzer® readout of purified and un-purified samples.	77
Figure 2.6. The Bioanalyzer® instrument and the chip	82
Figure 2.7. The wells of the High sensitivity DNA chip	83
Figure 2.8. Variations in the peak of 200-300bp between *FY* samples fragmented for 20 or 5 minutes	85
Figure 2.9. Variations in the peak of 200-300bp between *JK* samples fragmented for 12 or 8 minutes	86
Figure 2.10. The size distribution of an *ABO* sample fragmented at 6 minutes.	86
Figure 2.11.A A sample of the *JK* fragmented library (top graph) and the same library with adapter ligated (bottom graph)	89
Figure 2.11.B A sample of *FY* fragmented library (top graph) and the same library with adapter ligated (bottom graph)	89
Figure 2.12. The Pippin Prep™ Kit 2010 with Ethidium Bromide cassette	91
Figure 2.13. Two *JK* samples from the first experiment. (Top) experiment programmed to be size selected for broad range (peak around 200-250bp)	91

Figure 2.14. A purified size selected (PSS) *JK* sample assessed using the 2100 Bioanalyzer® instrument 98
Figure 2.15. Overview of the G810A analysis 103
Figure 2.16. Overview of the A588G analysis 104
Figure 3.1. Single long amplicon (generated by LR-PCR) covers the entire *FY* gene plus flanking regions (including the promoter region). 121
Figure 3.2. The *FY* amplicons produced by LR-PCR. 123
Figure 3.3. An electropherogram of a fragmented-purified *FY* DNA library. 125
Figure 3.4. Two electropherograms of the same *FY* amplicon library. 127
Figure 3.5. Two electropherograms of the same *FY* amplicon library, fragmented for 9 minutes and size selected by Pippin PrepTM. 128
Figure 3.6. three electropherograms of the same *FY* DNA library. 130
Figure 3.7 an electropherograms of size selected *FY* DNA library (by SPRIselect® reagent magnetic bead). 131
Figure 3.8 A summary report for a single sequencing run of *FY* library. 134
Figure 3.9 Mean Phred quality scores across all bases of *FY* samples in a single run. 138
Figure 3.10 The mean quality score of the *FY* sequences generated from a single run. 139
Figure 3.11 An IGV image illustrating the visualisation analysis of the sequencing data of a Fy (a+b+) sample with *FY*A*/*FY*B* genotype. 142
Figure 3.12 An example of a variant annotation report provided from the SeattleSeq Annotation 137 online tool of a Fy (a+b+) sample with *FY*A*/*FY*B* genotype. 144
Figure 4.1 The *JK* gene is illustrated with the three overlapping long amplicons generated by LR-PCR: 1 (11012bp), 2 (11053bp) and 3 (14665bp).These amplicons cover the entire *JK* gene and flanking regions. 165
Figure 4.2 Amplification of the entire *JK* gene using 3 amplicons produced by LR-PCR. 167
Figure 4.3 An electropherogram of a fragmented-purified *JK* DNA library (consists of a pool of 3 amplicons). 169
Figure 4.4 Two electropherograms of the same *JK* amplicon library. 171
Figure 4.5 Electropherograms of two different *JK* DNA libraries size-selected using either Pippin Prep™ or SPRIselect®. 172
Figure 4.6 A report of a representative single sequencing run of the *JK* library 174

Figure 4.7 Mean Phred quality scores across all bases of *JK* samples in a single run. 177
Figure 4.8 The mean quality score of the *JK* sequences generated from a single run. 178
Figure 4.9 An IGV output image showing visualisation analysis of the sequencing data of a Jk (a+b+) sample of *JK*A*/*JK*B* genotype. 180
Figure 4.10 An example of variant annotation report on a Jk (a+b+) sample (*JK*A*/*JK*B* genotype) obtained from the SeattleSeq Annotation 137 online tool. 181
Figure 4.11 Graphic illustration of the allele-specific SNPs forming reference *JK*A*, *JK*B* and *JK*Aw* allele sequences ('fingerprints'). 189
Figure 4.12 An IGV image illustrating the critical location of the 810G>A SNP in a *JK* sample. 193
Figure 4.13 Primer design for amplification of SNPs 810G>A and 838G>A. 194
Figure 4.14 The amplification of *JK* cDNA for splice site analysis 194
Figure 4.15 An electropherogram of 810G>A and 838G>A SNPs from Sanger sequencing of *JK*A/JK*B* cDNA. 195
Figure 4.16. An IGV image of the 810 G>A and 838 G>A SNPs and shows the splice site sequence between exon 8 and 9 of a *JK*A/JK*B* sample. 196
Figure 4.17 The primer pairs designed to cover exon 4 to exon 9 for 588A>G SNP analysis. 198
Figure 4.18 *JK* cDNA amplicon (864bp) containing SNP 588A>G. 199
Figure 4.19 *EcoRI* restriction enzyme analyses for positive insertions. 199
Figure 4.20 Electropherograms showing the association of SNP G588 with both *JK*B* and *JK*01W* alleles by Sanger sequencing of *JK*B/JK*01W* cDNA clones. 200
Figure 5.1 Four overlapping LR-PCR amplicons covering the entire *ABO* gene. 213
Figure 5.2 Four LR-PCR amplicons covering the whole *ABO* gene. 215
Figure 5.3 An electropherogram of a purified fragmented *ABO* DNA library (consists of a pool of four *ABO* amplicons). 217
Figure 5.4 Two electropherograms of an *ABO* sample library. 219
Figure 5.5 An electropherogram of a size-selected *ABO* sequencing library by SPRIselect®. 220
Figure 5.6 A summary report for a single *ABO* library sequencing run. 222
Figure 5.7 The Phred quality score across all bases of a single representative *ABO* sample. 224
Figure 5.8 The mean quality score of *ABO* sequences generated from a single sample. 226
Figure 5.9 An IGV image illustrating the full coverage of the sequencing data from 227

one *ABO* sample.
Figure 5.10 IGV visualisation analysis of *ABO* sequencing data of a single sample. 229
Figure 5.11 illustrations regarding the *ABO* reference sequence. 230
Figure 5.12 IGV image of an *ABO* sample of phenotype O where coverage of areas by amplicon 1A failed. 233
Figure 5.13 A section of the report generated by the Ion Reporter™ showing genotyping analysis of an *ABO* sample with B phenotype and *B101(ABO*B.01)*/ *O01(ABO*O.01.01*) genotype. 235
Figure 5.14 Sanger sequencing of a number of SNPs to confirm NGS data of one sample. 256
Figure 6.1 Different *JK*A* alleles based on the intronic polymorphism patterns. 284

List of Tables

Table 1.1 Blood Group Systems	5
Table 1.2 Basic concept of ABO system with the relation between antigens and antibodies.	15
Table 1.3 Examples of different NGS platforms with their sequencing chemistries. (Adapted from (Hodkinson and Grice, 2015)	48
Table 1.4 Updated chips v2 of Ion PGMTM in terms of capacity and run time according to read length	54
Table 2.1. LR-PCR forward (F) and reverse (R) primers for *FY* amplification.	68
Table 2.2. LR-PCR forward (F) and reverse (R) primers for *JK* amplifications	68
Table 2.3. LR-PCR of forward (F) and reverse (R) primers for *ABO* amplifications.	69
Table 2.4. Optimised thermocycling conditions for *FY*.	71
Table 2.5. Optimised thermocycling conditions for *JK*.	71
Table 2.6. Optimised thermocycling conditions for *ABO* amplicons 2, 3 and 4.	72
Table 2.7. Optimised thermocycling conditions for *ABO* amplicons 1A and 1B.	72
Table 2.8.A. The size selection range programme for *FY* samples.	94
Table 2.8.B. The size selection range programme for *JK* samples.	94
Table 2.9. The primers designed to cover the gene region within exons 8 and 9 of *JK* for the analysis of G810A (exon 8 near splice region).	106
Table 2.10. The primers designed to cover the gene region within exons 4 to 9 of *JK* for the analysis of A588G in exon 7 with *JK*01W*.	106
Table 2.11. Thermocycling conditions for cDNA amplification, specific for the analysis of the *JK* (G810A on exon 8 near splice region).	109
Table 2.12. Optimised thermocycling conditions for cDNA amplification, specific for analysis of *JK* A588G in exon 7.	110
Table 2.13 Primers designed and used to confirm number of *ABO* SNPs from NGS data.	115
Table 2.14 Thermocycling conditions for amplification of the primers in Table 2.13 used for the confirmation step.	116
Table 3.1. The serology information of the 53 blood samples provided from	122

National Health Service Blood and Transplant (NHSBT;Filton, Bristol UK).
Table 3.2. A summary of the Ion PGMTM sequence run output of the 53 *FY* samples, collectively processed in 3 runs. 133
Table 3.3 Phred quality score and the base call accuracy. 137
Table 3.4 Summary of the sequencing quality of the 3 *FY* NGS runs. 140
Table 3.5 NGS genotyping of 53 samples of differing Fy phenotypes. 147
Table 3.6 The list of SNPs in the intron of the *FY* gene and flanking regions. 148
Table 3.7 NGS genotyping of 53 *FY* samples, 43 of which are of known phenotype. 149
Table 4.1 Serology information on the 67 blood samples provided by the National Health Service Blood and Transplant (NHSBT; Filton, Bristol UK). 166
Table 4.2 A summary of the Ion PGMTM sequence report of the 67 *JK* samples 173
Table 4.3 Summary of the sequencing quality of the four *JK* NGS runs. 176
Table 4.4. NGS genotyping of 67 samples of differing JK phenotypes. 184
Table 4.5 NGS of 67 different *JK* samples, 59 of which were of known Jk phenotype. 187
Table 5.1 The serology information of all 47 blood samples provided by the National Health Service Blood and Transplant (NHSBT; Filton, Bristol UK). 214
Table 5.2 A summary of the Ion PGMTM sequence output for the 47 *ABO* samples. 221
Table 5.3 Summary of the sequencing quality of the three *ABO* NGS experiments. 225
Table 5.4 The differences in polymorphisms between the reference sequence in hg19 and the agreed consensus sequence (*ABO*A1.01*) from the NCBI. 231
Table 5.5 The SNPs found around the 1A primer pair binding sites in examples of *ABO* samples with different phenotype. 233
Table 5.6 NGS genotyping of 47 samples of different ABO phenotypes. 239
Table 5.7 NGS genotyping of 47 ABO samples (all provided with serological phenotype). 247
Table 5.8 NGS genotyping of homozygous *O* allele samples. 251
Table 5.9 The analysis of minisattelite repeats 3.8kb upstream of the start codon of *ABO*. 254

Abbreviations

A	Adenine
aa	amino acid
ACKR1	Atypical Chemokine Receptor
BAM	binary alignment/map
BGG	blood group genotyping
BGMUT	Blood Group Antigen Gene Mutation Database
bp	base pair
BR	Broad range
C	cytosine
cfDNA	cell-free DNA
DAT	direct antiglobulin test
ddNTPs	deoxynucleotide triphosphates
DHTRs	delayed HTRs
DNA	deoxyribonucleic acid dideoxynucleotides triphosphates
dNTPs	deoxynucleotide triphosphates
EDTA	ethylenediaminetetraacetate
FHM	foeto-maternal haemorrhage
FUC	Fucose
G	guanine
Gal	galactosyl
GalNAc	*N*-acetyl- D-galactosamine
gDNA	Genomic DNA
GDP-L-fucose	guanosine diphosphofucose
GTA	3-α-*N*-acetylgalactosaminyltransferase
GTB	3-α-galactosyltransferase
HDFN	haemolytic disease of the foetus and newborn
HEA	human erythrocyte antigen
hg	human genome
HGNC	HUGO Gene Nomenclature Committee

HGP	Human Genome Project
HS	High sensitivity
HTR	haemolytic transfusion reaction
IAT	indirect antiglobulin test
IgG	immunoglobulin G
IgM	immunoglobulin M
IGV	Integrative Genome Viewer
Ion PGM™	Ion Torrent Personal Genome Machine™
ISBT	International Society of Blood Transfusion
ISPs	ion sphere particles
Kb	kilobases
LR-PCR	long-range polymerase chain reaction
mRNA	messenger RNA
NCBI	National Centre for Biotechnology Information
NGS	next-generation sequencing
NHGRI	National Human Genome Research Institute
NHSBT	National Health Service Blood and Transplant
ONT	Oxford Nanopore Technology
PCR	polymerase chain reaction
RBCs	red blood cells
RFLP-PCR	restriction fragment length polymorphism
RNA	ribonucleic acid
RT-PCR	real time PCR
SBT	sequencing-based typing
SCD	sickle cell disease
SNPs	Single nucleotide polymorphisms
SNV	single nucleotide variant
SSP-PCR	sequence specific primer-PCR
T	Thymine
TAPS	N-Tris(hydroxymethyl)methyl-3-aminopropanesulfonic acid
T_m	melting temperature
VCF	variant call format
WGS	whole genome sequencing

WGS	whole genome sequencing
α2FUCT1	2-α-L-fucosyltransferase
β-ME	β-mercaptoethanol

Amino Acids

Amino acids	Three-letter code
Alanine	Ala
Arginine	Arg
Asparagine	Asn
Aspartic acid	Asp
Cysteine	Cys
Glutamic acid	Glu
Glutamine	Gln
Glycine	Gly
Histidine	His
Isoleucine	Ile
Leucine	Leu
Lysine	Lys
Methionine	Met
Phenylalanine	Phe
Proline	Pro
Serine	Ser
Threonine	Thr
Tryptophan	Trp
Tyrosine	Tyr
Valine	Val

Chapter 1

Introduction

1.1 Introduction to blood groups

Blood groups were first discovered by Karl Landsteiner in early 20th century (particularly 1901). He found that plasma samples of a number of individuals showed agglutination when exposed to red blood cells (RBCs) of others but not with their own RBCs. Subsequent investigations led to the identification of the three ABO blood groups (A, B and O) for the first blood group system, ABO. Subsequently, in the following year, a fourth group (AB) for the ABO system was described by Decastello and Sturli (Landsteiner, 1961, Levine, 1961, Owen, 2000, Watkins, 2001, Schwarz and Dorner, 2003, Giangrande, 2000).

Each blood group is characterised by specific antigens, which can be defined as inherited polymorphic markers present on the external membrane of RBCs and are detected by specific antibodies. The blood group antigens can elicit an immune response, which is marked by the synthesis of corresponding antibodies that occur naturally, as a result of exposure to antigens in the environment or alloimmunisation of those who lack the antigens that are acquired through blood transfusion or pregnancy. These specific alloantibodies are utilised to serologically detect blood group antigens (Reid et al., 2012, Daniels, 2013). These antigens are either protein or carbohydrate molecules attached to protein or lipid of the RBC membrane forming protein, glycoprotein, or glycolipid structures (Reid and Mohandas, 2004, Daniels, 2005). The various blood group antigens are encoded or controlled by either a single gene, such as the JK blood group system, or two or three genetically discrete genes that exist as a cluster of closely linked homologous genes, as in the case of the RH and MNS blood group systems, respectively (Daniels, 2013). The blood group genes either encode for the enzyme that

transfers the carbohydrate determinants to the membrane glycoprotein or glycolipid, as is the case with the ABO antigen formation, or directly encode for the amino acid (aa) sequence in the protein determinates blood group antigens, such as those of JK (Reid and Mohandas, 2004). These genes have been mapped to diverse chromosomes in the human genome (Storry and Olsson, 2004). The International Society of Blood Transfusion (ISBT) has recognised more than 300 blood group antigens that are categorised into 36 blood group systems (Table 1.1) (ISBT, 2016).

1.2 Blood group terminology

In 1980, the ISBT set up a working party on Red Cell Immunogenetics and Blood Group Terminology, which later became a committee, in order to establish a uniform nomenclature that is based on the genetic bases of the antigens and is easily readable by machine, software and eye. Blood group antigens (authenticated) are classified into four categories: systems, collections, low incidence antigens (700) and high incidence antigens (901) (Daniels et al., 2004). A blood group system consists of one or more antigens that are controlled by a single gene or by two or three closely linked homologous genes. Collections represent sets of antigens that are serologically, biochemically or genetically related but do not fulfil the requirements for inclusion into a blood group system, which is suggested because lack of information regarding the gene identity. Nevertheless, upon identification of the gene, a collection might subsequently become recognised as a system and become an obsolete collection. Examples are the Cromer collection (202) that became a blood group system (021) (Daniels, 2013, Reid et al., 2012) Table 1.1) and the 212 Vel (212) collection, which was declared obsolete and categorised as the VEL (034) blood group system after identification of the relevant gene (Storry et al., 2013, Ballif et al., 2013, Cvejic et al., 2013). The antigens that have an incidence of less than 1% in most populations and are

not included in the system or collection criteria, fit into the 700 series. The 901 series, previously known as 900, consists of high-frequency antigens, with incidence greater than 90% in populations, but do not belong to system or collections. Similarly, those antigens might be included in a blood group system if the pertinent criteria are fulfilled, as seen in the case of Lan (901 high-incidence antigen) that was declared as the Lan blood group system after the identification of the corresponding gene (*ABCB6*) in 2012 (Helias et al., 2012, Storry et al., 2014).

With regard to the ISBT terminology of the blood group antigens, each antigen belonging to a system is denoted and identified by six digits, the first three of which represent the blood group system, for instance 008 for the FY system, whereas the last three identify the antigen. For example, 008001 denotes the first antigen of the FY system (Fy^a). Alternatively, the system alphabetical symbol can be used instead of the numerical ones, followed by the antigen number with (FY001) or without the sinistral zeros (FY1) to represent the antigen (Fy^a). Phenotypes are represented by the system symbol followed by a colon then by a list of the antigens present and separated by a comma, while absent antigens marked by a minus (for example FY: -1, 2 for Fy (a-b+)). Alleles are denoted by the system symbol followed by an asterisk then the antigen number or letter, all in italics; for example, *FY*01* or *FY*A* encodes for Fy^a. Genotypes are denoted by the system symbol followed an asterisk then the alleles or haplotypes are separated by a slash, all in italics; for example, *FY*01/02* or *FY*A/B*. Null or amorph alleles are represented by zero; for example, *KEL*0* or letter N as for example *FY*01N.01* is responsible for the Fy null phenotype. The gene is denoted by the italicised symbol of the ISBT blood group system, which is mostly used instead of that defined by the HUGO Gene Nomenclature Committee (HGNC), in which the names reflect the function of the gene product. The nomenclature for the antigens in collections, 700 and 901 is similar to that in the system where the system digits are replaced by the

collection number or 700 and 901, respectively.

The establishment of the numerical terminology was mainly for the purpose of computer storage of information on the blood group antigens and genetic classification. However, the routine use of such terminology might not be suitable, which has led to emergent of alternative approved terminology for blood group antigens and phenotype that are popular and more user-friendly. Examples are Fy^a instead of FY001 and Fy (a+b-) instead of FY:1,-2 for the antigen and phenotype, respectively, see (see the ISBT website for more examples) (ISBT, 2016, Daniels et al., 2004, Daniels et al., 2007, Storry et al., 2011, Daniels, 2013). The 36 blood group systems are listed in Table 1.1, with the system name, the ISBT symbol, ISBT number, number of antigens per system, the gene(s) encoding for the antigens and the chromosome (ISBT, 2016).

Table 1.1 Blood Group Systems. * When in italic, ISBT gene name. ** Gene name according to HUGO Gene Nomenclature Committee (HGNC) (HGNC, 2016). Adapted from (ISBT, 2016).

ISBT Number	System name	ISBT Symbol*	No. of Antigens	Gene name (**)	Chromosome
001	ABO	ABO	4	*ABO*	9
002	MNS	MNS	48	*GYPA,GYPB,GYPE (3)*	4
003	P1PK	P1PK	3	*A4GALT*	22
004	Rh	RH	54	*RHD,RHC (2)*	1
005	Lutheran	LU	22	*BCAM*	19
006	Kell	KEL	35	*KEL*	7
007	Lewis	LE	6	*FUT3*	19
008	Duffy	FY	5	*ACKR1*	1
009	Kidd	JK	3	*SLC14A1*	18
010	Diego	DI	22	*SLC4A1*	17
011	Yt	YT	2	*ACHE*	7
012	Xg	XG	2	*XG, CD99 (2)*	22
013	Scianna	SC	7	*ERMAP*	1
014	Dombrock	DO	10	*ART4*	12
015	Colton	CO	4	*AQP1*	7
016	Landsteiner-Wiener	LW	3	*ICAM4*	19
017	Chido/Rodgers	CH/RG	9	*C4A, C4B (2)*	6
018	H	H	1	*FUT1*	19
019	Kx	XK	1	*XK*	21
020	Gerbich	GE	11	*GYPC*	2
021	Cromer	CROM	18	*CD55*	1
022	Knops	KN	9	*CR1*	1
023	Indian	IN	4	*CD44*	11
024	OK	OK	3	*BSG*	19
025	Raph	RAPH	1	*CD151*	11
026	John Milton Hagen	JMH	6	*SEMA7A*	15
027	I	I	1	*GCNT2*	6
028	Globoside	GLOB	2	*B3GALT3*	3
029	Gill	GIL	1	*AQP3*	9
030	RHAG	RHAG	4	*RHAG*	6
031	Forsman	FORS	1	*GBGT1*	9

032	Junior	JR	1	*ABCG2*	4
033	Lan	LAN	1	*ABCB6*	2
034	VEL	Vel	1	*SMIM1*	1
035	CD59	CD59	1	*CD59*	11
036	Augustine	AUG	2	*SLC29A1*	6

1.3 The molecular basis of blood group polymorphisms

The cloning of the blood group genes allows the study of the genetic bases of blood groups, revealing polymorphisms and variations. This has paved the way for better understanding and investigation of the molecular background of blood group antigens. The cloning of blood group genes was first performed in 1986, when the blood group gene *GYPA* encoding the MN blood group antigens (of the MNS system) was cloned (Siebert and Fukuda, 1986). Subsequently, more genes were cloned to extend the investigation at the molecular level, such as *ABO* and *RH* in between 1990 and 1992 (Yamamoto et al., 1990b, Avent et al., 1990, Le van Kim et al., 1992). Consequently, with the advances made in DNA analysis technology, such as sequencing techniques, more data regarding the molecular basis of most of the clinical blood group genes, such as polymorphisms, were acquired and suggested to be known (Daniels, 2013). The polymorphism can be defined as a character that is present in two or more different forms in a single population (Daniels, 2005). The expression of the diverse blood group antigens involves various polymorphisms in the genes at the molecular level. In fact, a number of genetic mechanisms account for polymorphisms, for instance single nucleotide polymorphisms (SNPs) and gene deletion. SNPs, which can be defined as a substitution of nucleotides in the gene, are the most common cause of the genetic diversity among blood group systems, if not the entire human genome. In fact, around 1.42 million SNPs have been recorded in the human genome map (Sachidanandam et al.,

2001), and this number increased significantly to almost 15 million SNPs as reported by the 1000 genome project in 2010 (Genomes Project et al., 2010) and the NCBI database (www.ncbi.nlm.nih.gov/snp).

Polymorphisms can exist at various points in the gene, such as exons, introns, in splice regions (between exons and introns) and even in regulatory regions. With regard to SNPs occurring in exons (the coding region of the genes), they could be silent, whereby no apparent change may occur in the aa sequence or changes may be noted in the encoded product, i.e. the protein or the enzyme (glycosyltransferase), manifesting as missense or non-sense mutations. The missense SNPs in exons that lead to aa substitutions in the gene product account for the majority of blood group polymorphisms and antigen diversity (Storry and Olsson, 2004, Daniels, 2005). Nonsense SNPs are reported to cause substitution of the aa to a premature stop codon, resulting in a truncated gene product that might affect antigen expression. Such SNPs were reported in a previous study (Rios et al., 2000) which indicated the presence of various nonsense mutations in the exon 2 of the *FY* gene in three unrelated individuals belonging to different ethnicities with phenotype Fy (a-b-). The mutation changes the tryptophan to a premature stop codon, which prevents the expression of the Fy antigen on the red cell membrane and as suggested from all tissues. The mutations occurred at different points in each individual that is, G>A mutation at nucleotides 287, 408 in *FY*A* background and 407 in *FY*B* background (Rios et al., 2000). In addition to abolishing antigen expression, SNPs may weaken the antigenicity, reduce the number of antigen copies expressed, as seen in the Fy^x phenotype, reducing the expression levels of Fy antigens, particularly Fy^b, on the RBC surface due to missense SNPs encoding aa change mainly, Arg89Cys along with the Ala100Thr (Tournamille et al., 1998, Olsson et al., 1998, Gassner et al., 2000). The emergence of polymorphisms is believed to occur as the result of exposure to pathogens, and therefore, the individuals' blood group

antigens evolve to tolerate such pathogens. This can be seen in the case of individuals with the null Fy phenotype with no expression of the antigen which is suggested to be a receptor for the malaria parasite (Miller et al., 1975, Miller et al., 1976); such individuals are most likely resistant to the malarial parasite (to be discussed in FY section 1.6.1). Polymorphisms in introns, especially, at splice sites, might disrupt antigen expression. Splice site sequences are important for exon fusion after the removal of the introns in the mechanism of splicing during the formation of messenger RNA (mRNA) in the transcription process. Mutations that affect these splice sequences, particularly the conserved GT in the intron of the 5′-donor splice site or the AG at -1 and -2 of the 3′-acceptor splice site (Berget, 1995, Zhang et al., 2003) might affect this process, leading to exon skipping and altered antigen expression. Examples of this effect can be seen in two studies investigating the null phenotypes JK and KEL. With regard to the JK_{null} phenotype, the analysis of Polynesian individuals with the phenotype Jk(a-b-) showed a SNP in the invariant G in the 3′-acceptor splice site of intron 5 to A that resulted in exon 6 skipping, thereby affecting antigen expression (Irshaid et al., 2000). Similarly, the K_0 phenotype, in which K antigen is absent from the RBC membrane, was attributed to a SNP in the invariant GT>AT sequence at the end of the 5′-donor splice sequence in intron 3 (Lee et al., 2001). Mutations have been reported to occur in regulatory regions that might alter antigen expression. A common example would be as in African individuals with the Fy(a-b-) phenotype, in which the Fy antigen expression is abolished from the RBCs as a result of a SNP in the binding site of the erythroid transcription factor GATA-1 in the promoter region of the *FY*B* background (Tournamille et al., 1995). This is, as mentioned before, a manifestation of the suggested evolution to avoid the pathogens since the FY antigen is a receptor for the malarial parasite and, especially *Plasmodium vivax,* which implies that those with such mutation are protected against infestation of *P. vivax* (Miller et al., 1976). Those with

null phenotypes, particularly rare ones, are vulnerable to immunisation through blood transfusion, and consequently, if immunised, the availability of matched blood would be difficult due to its rarity (Storry and Olsson, 2004). An example of which is the rare Rh_{null} phenotype, characterised by the lack of expressions of all Rh antigens, that may produce anti-Rh29 (which reacts with RBCs of all Rh phenotypes except Rh_{null}), if immunised (Daniels, 2013).

In addition to SNPs, there are other genetic mutation mechanisms that account for polymorphisms of the blood groups. Deletion or insertion of nucleotides or the entire gene deletion can also affect the blood group antigen expression. The deletion of the entire *RHD* gene is a common reason for the absence of the expression of the D antigen leading to D-negative phenotype in Caucasians (Wagner and Flegel, 2000).

Deletion of nucleotides has been reported to cause shift in the open reading frame of the sequence, thereby affecting the encoded amino acid (aa) after the deletion and altering antigen expression (Daniels, 2005). This is illustrated in two diverse single nucleotide deletions in exons 6 and 7 of the *A101* allele, accounting for the *O01* and *A201* alleles, respectively (Yamamoto et al., 1990a, Yamamoto et al., 1992). Furthermore, sequence duplication can also account for inactivation and loss of expression of antigens, as seen in Africans carrying an inactive *RHD* gene (*RHDψ*), pseudogene, with 37-bp duplication split between intron 3 and exon 4; this results in a reading-frame shift and premature stop codon, along with a nonsense mutation (Tyr269stop) in exon 6 disturbing the RhD antigen expression on RBCs (Singleton et al., 2000). Intergenic recombination between homologous genes might alter or eliminate the expression of antigens and lead to various phenotypes. The Rh blood group system has been reported in cases of hybrid gene or crossover between *RHD* and *RHCE* in sharing sequences; this can result in a hybrid gene, *RH (D-CE-D*) that might abolish the expression of the D antigen in Africans (Daniels, 2005). Another example of the hybrid gene is the one

possibly responsible for the partial D phenotype (D^{VI} type I); this phenotype is encoded by a hybrid *RH (D-CE-D*) gene, in which exons 4 and 5 of the *RHD* gene are replaced by those of the *RHcE* allele (Avent et al., 1997).

1.4 The immunogenicity of blood groups

A number of blood group antibodies are considered clinically significant, leading to adverse reactions by damaging cells, especially RBCs, which carry the corresponding antigens. These antibodies may be naturally existing; for example, antibodies, mostly IgM, against the absent antigens of the ABO system are naturally present in blood. On the other hand, alloimmunisation with immunoglobulin G (IgG) antibodies occurs as a consequence of the exposure to blood group antigens absent in an individual, which generally occurs via transfusion or foeto-maternal haemorrhage during pregnancy. These antibodies result in various adverse effects, mainly haemolytic transfusion reaction (HTR) and haemolytic disease of fetus and newborn (HDFN) (Daniels, 2013), which are briefly described in the next subsections.

1.4.1 Haemolytic transfusion reaction (HTR)

The transfusion of mismatched blood units, particularly red cell units, may lead to the destruction of the transfused red cells (i.e. an HTR); this reaction can be mild or fatal. Intravascular HTR is characterised by rapid haemolysis, in which the circulatory RBCs are damaged within 10 minutes. This antibody-mediated haemolysis is caused by IgM antibodies that can activate the complement classical pathway, forming the membrane attack complex and puncturing the RBC membrane. As a result, free haemoglobin is released into the plasma, followed by clinical signs, such as haemoglobinemia,

haemoglobinuria, hypotension, chills, shock, and more serious complications of disseminated intravascular coagulation and renal failure. ABO antibodies, which are mainly IgM, are commonly the cause for intravascular HTR (Poole and Daniels, 2007, Dean, 2005).

HTR can be extravascular, where the IgG antibodies are involved. This mostly occurs in the case of mismatch with clinically significant antibodies other than those of the ABO system. IgG antibodies rarely activate the complement to initiate the haemolytic process, but rather coat the RBCs which then bind the Fc receptors on macrophages in the spleen or liver and destroy them by phagocytosis. The clinical manifestations are similar to those of intravascular HTR but less severe. The extravascular HTR can be immediate (few hours) or delayed (few days, typically 5–7 days); the immediate reaction occurs when the antibody level is serologically detectable, whereas the latter occurs in cases with low levels of antibodies (often too low to be detected by serology). Examples of antibodies causing extravascular HTR are those of the Rh and JK systems (Poole and Daniels, 2007).

1.4.2 Haemolytic disease of fetus and newborn (HDFN)

HDFN occurs as a consequence of immunisation of an antigen-negative mother exposed to antigen-positive foetal RBCs (inherited from father), usually after foeto-maternal haemorrhage (FHM). As a result, the immune system of the mother is sensitised and forms antibodies, which are initially IgM that do not to cross the placenta. In subsequent pregnancies, the maternal antibodies are IgG (mostly IgG1 and IgG3), which are capable of crossing the placenta to enter the foetal circulation, acting against the foetal antigen on RBCs and coating the RBCs, leading to their destruction by macrophages in the spleen (Urbaniak and Greiss, 2000, Dean, 2005). This might lead to complications, such as jaundice, or pose risk to the foetus' life. The antibodies of various blood group

systems, such as anti-D,-c, -K, Fy and Jk, are able to cause HDFN at different levels of severity although anti-D is known to be the most common cause (Basu et al., 2011, Poole and Daniels, 2007). Those of ABO, FY and JK will be discussed later (sections 1.5.5, 1.6.3 and 1.7.3).

A preventative management of HDFN, particularly that of anti-D, has been applied by the administration of IgG anti-D prophylaxis to unsensitised D-negative mothers. This likely reduces the risk of immunisation by coating the foetal RBCs leaked to the mothers' circulation and thus rapidly destroying and removing these cells by macrophages in the spleen before triggering the mother's immune response of alloimmunisation with anti-D (Kumpel, 2008).

1.5 ABO blood group system

In 1900-1901, the ABO system (001) was the first blood group system discovered by Karl Landsteiner (Landsteiner, 1961; Levine, 1961; Owen, 2000). In transfusion medicine, it is considered to be one of the most important blood group systems. This is because in contrast to other blood groups systems, such as the Rh system, the antibodies against A or B antigens are naturally present in the blood circulation of those with the absent antigens (Yamamoto, 2004, Storry and Olsson, 2009), thereby definitively causing a critical transfusion reaction in the case of mismatched blood units. The ABO system consists of A, B, AB and A1 antigens, which are expressed on the red cells and on the surface of other types of cells and secretions. Accordingly, it can be also referred to as a histo-blood group system rather than a blood group system alone (Yamamoto, 2004). As a result, matching for the ABO group is critical for both safe blood

transfusion and safe transplantation of cells, tissue and organs. Moreover, it can be used in forensic investigation for the detection of saliva, blood and hair traces found in crime scenes and used in suspect exclusion. Although the main ABO groups are A, B, AB and O, numerous subgroups encoded by a considerable number of alleles (381 alleles) with various polymorphisms (dbRBC, 2016) have also been reported. A number of these polymorphisms alter enzyme activity, resulting in various degrees of agglutination and patterns (Chester and Olsson, 2001). Pathogens and diseases may affect the expression of A and B antigens; for example, in cancerous tumours, glycosylation of the cell surface proteins is altered, which could be due to the down-regulation of the glycosyltransferase and thus affect the expression of A and B antigens (Dabelsteen and Gao, 2005, Yamamoto, 2004).

1.5.1 Discovery

The ABO blood group system was the first blood group system discovered, and therefore, it is marked with the number 001 as per the official ISBT terminology system (Daniels et al., 2004). In 1901, Landsteiner observed and described agglutination of individuals' RBCs with liquid components (sera) of some of his colleagues but not with those of others. This phenomenon was later postulated to be due to the existence of naturally occurring antibodies (anti-A and anti-B) against the absent antigens in individuals; these antibodies probably occur as a result of natural immunisation by stimulation of substances, such as bacteria, food and pollen. In other words, anti-A and anti-B antibodies are absent in the serum of individuals who show expression of the respective antigens in their own red cells, whereas both antibodies are present in those with group O, who do not express any of the ABO antigens on their red cell surfaces (Watkins, 2001) (Table 1.2). Subsequently, according to the agglutination pattern, the

ABO system was categorised by Landsteiner into three main groups, A, B, and O (originally called C), with O being derived from the German word “ohne”, which means “without” (Landsteiner, 1961, Levine, 1961, Owen, 2000, Watkins, 2001). Subsequently, one more group (AB) characterised by expression of both antigens and absence of both antibodies was described in the following year by Decastello and Sturli reviewd by (Watkins, 2001), thus making four blood groups in the ABO system and four basic phenotypes (A, B, AB and O) (Yamamoto, 2004) (Table 1.2).

Table 1.2 Basic concept of ABO system with the relation between antigens and antibodies.

ABO Blood Group	Red Cell Antigen	Antibodies in Serum
A	**A**	**Anti-B**
B	**B**	**Anti-A**
O	**None**	**Anti-A,B**
AB	**A and B**	**None**

1.5.2 Biosynthesis of ABO blood group antigens

The A and B antigens along with the H antigens are carbohydrate determinants on glycoproteins (mainly) or glycolipids on red cells and other tissues (Yamamoto, 2004). The synthesis process can be described as a sequential addition of monosaccharides into oligosaccharide chains with polypeptides forming glycoproteins or with ceramide to form glycolipids (Clausen et al., 1994). The process is initiated by the action of the glycosyltransferase enzymes, the product of the *ABO* system genes, in transferring the specific immunodominant sugar of A or B antigens from the nucleotide substrate (donor) to the acceptor substrate (the H antigens), which is the main carbohydrate foundation for the formation of the A and B antigens (Storry and Olsson, 2009).

The synthesis of the ABO antigens is dependent on the H antigens, which is the main foundation upon which the structure of A and B antigens are built (Morgan and Watkins, 1948). H antigen is inherited by two genes, *FUT1* and *FUT2*, located on chromosome 19. *FUT1* encodes for 2-α-L-fucosyltransferase (α2FUCT1), which is responsible for the production of H antigen on precursor type 2 that lies on mesodermal cells, including RBCs, whereas *FUT2* is responsible for H antigen synthesis in secretions mainly on precursor type 1 with 2-α-L-fucosyltransferase (α2FUCT2) enzyme. Briefly, these enzymes catalyse the transfer of the Fucose (FUC) from guanosine diphosphofucose

(GDP-L-fucose) donor to the terminal galactose of one of the precursor types. These structures act as an acceptor substrate for the carbohydrate determinants transferred by the A and B glycotransferase, GTA and GTB, respectively, which are the products of the *ABO* gene to form A and B antigens (Daniels, 2013).

The ABO blood group is occasionally called the ABO histo-blood group due to its widespread distribution among the tissues, depending on the H antigen and the precursor type (Storry and Olsson, 2009). The antigens found on the type 2 precursor and encoded by *FUT1* account for the main ABO antigens present on red cells, whereas type 1 precursor along with *FUT2* are mainly the holders for the ABO antigens in secretions, for example plasma (Mollicone et al., 1995). Consequently, the A and B antigens are absent from red cells and secretions in individuals who lack H antigen (silenced *FUT1* and *2*); this phenotype is referred to as the Bombay phenotype (Oh) (Storry and Olsson, 2009, Moores et al., 1975, Yunis et al., 1969).

Blood group A is formed as a result of the act of the *A* allele that encodes for an enzyme 3-α-*N*-acetylgalactosaminyltransferase (GTA), which catalyses the transfer of *N*-acetyl-D-galactosamine (GalNAc), the immune dominant sugar of A antigen from the donor substrate uridinediphosphate (UDP)-N-acetylgalactosamine to the fucosylated galactosyl (Gal) residue of the H antigen (H precursor) Fuc α (1-2)Galβ-GlcNAcβ-R. On the other hand, individuals with blood group B possess the *B* allele that encodes for the 3-α-galactosyltransferase (GTB), which catalyses the transfer of galactose from UDP-galactose to the H precursor via a process similar to the one described for the A group (Daniels, 2005, Hosoi, 2008) see (Figure 1.1). With respect to group O, however, its alleles lack the production of active enzyme, leaving the H antigens unaltered, thereby suggesting that those with blood group O illustrate the highest amount of H antigens (Daniels, 2005, Yamamoto et al., 1993). AB individuals inherit both *A* and *B* alleles with GTA and GTB activities.

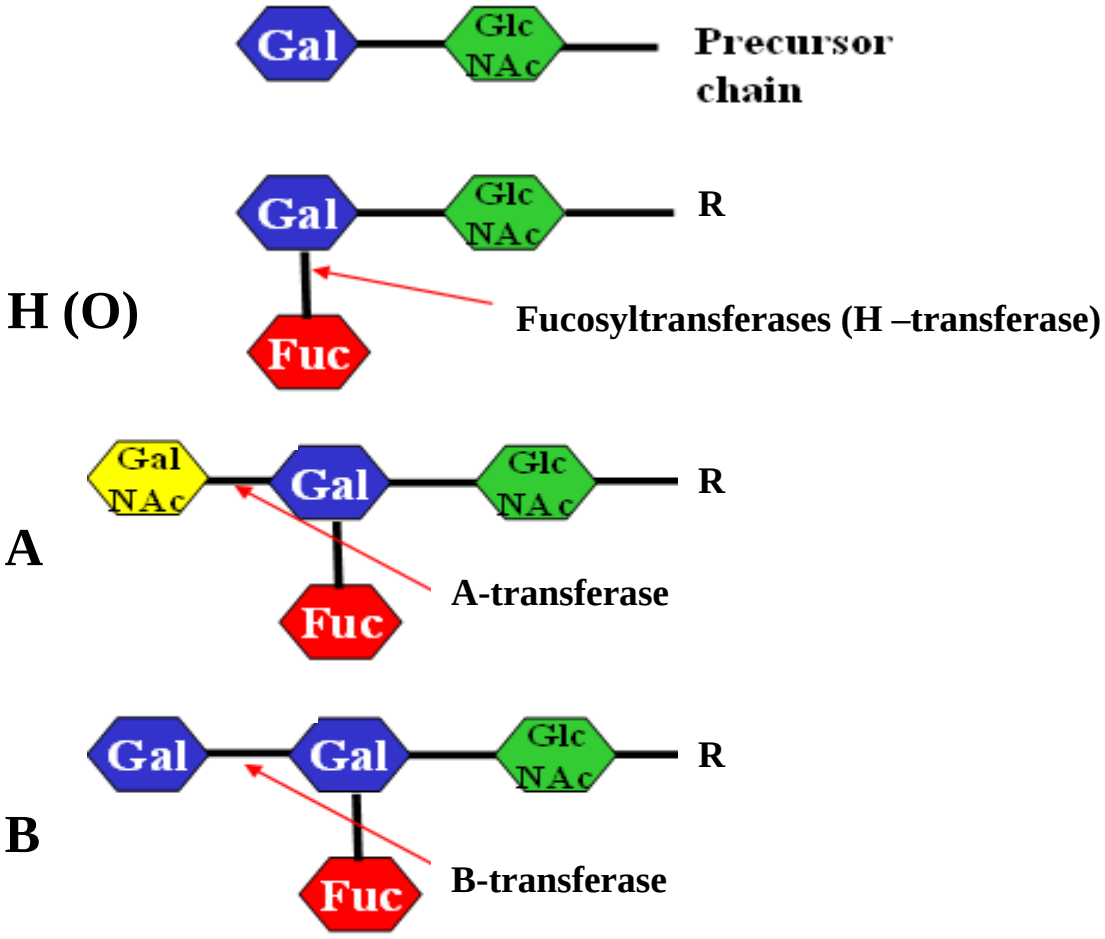


Figure 1.1 A brief diagram of A, B and H antigen (also O group) oligosaccharides structure biosynthesis.

The process of the synthesis of the group A and B antigens from H antigen is shown, by A-transferase and B-transferase enzymes, the products of *A* and *B* alleles. R, remainder of molecule; GlcNAc, N-acetylglucosamine; Gal, galactose; Fuc, fucose; GalNAc, N-acetylgalactosamine; UDP, uridine diphosphate. Adapted from (Cooper and Group, 2003).

1.5.3 Inheritance and molecular genetics

The ABO antigens, carbohydrate antigens, are encoded indirectly by the *ABO* gene as the gene encodes for the glycosyltransferase enzymes, which are involved in the synthesis of the antigens (Green, 1989, Watkins, 2001). The ABO blood group system are recognised as inherited, with the A and B antigens being inherited codominantly and O being inherited recessively (Storry and Olsson, 2009, Yamamoto, 2004).

The ABO glycosyltransferases are transmembrane proteins of the Golgi apparatus that consist of short *N*-terminal domain (cytoplasmic) and large C-terminal catalytic domain (Figure 1.2) (Paulson and Colley, 1989). They also show to contain two domains that recognise the acceptor substrate (H antigen) and the donor substrate (UDP-Gal or UDP-GalNAc) (Patenaude et al., 2002). Therefore, the effect of the genetic variations and polymorphisms would further increase the complexity of the expression due to the sequential actions of different gene (ABO and H) products that are enzymes involved in the synthesis of the ABO antigens.

The ABO cDNA encodes for GTA was cloned by Yamamoto (Yamamoto et al., 1990b). The *ABO* gene is located on the long arm of chromosome 9 (9q34) (Bennett et al., 1995), consisting of 7 exons spanning about 20 kilobase pairs (kbp) that encode for glycotransferase of 354 aa. Exons 6 and 7 hold the majority of the coding sequence, accounting for 77% of the enzyme, particularly the catalytic domain (Yamamoto et al., 1995, Bennett et al., 1995, Daniels, 2013) (Figure 1.3).

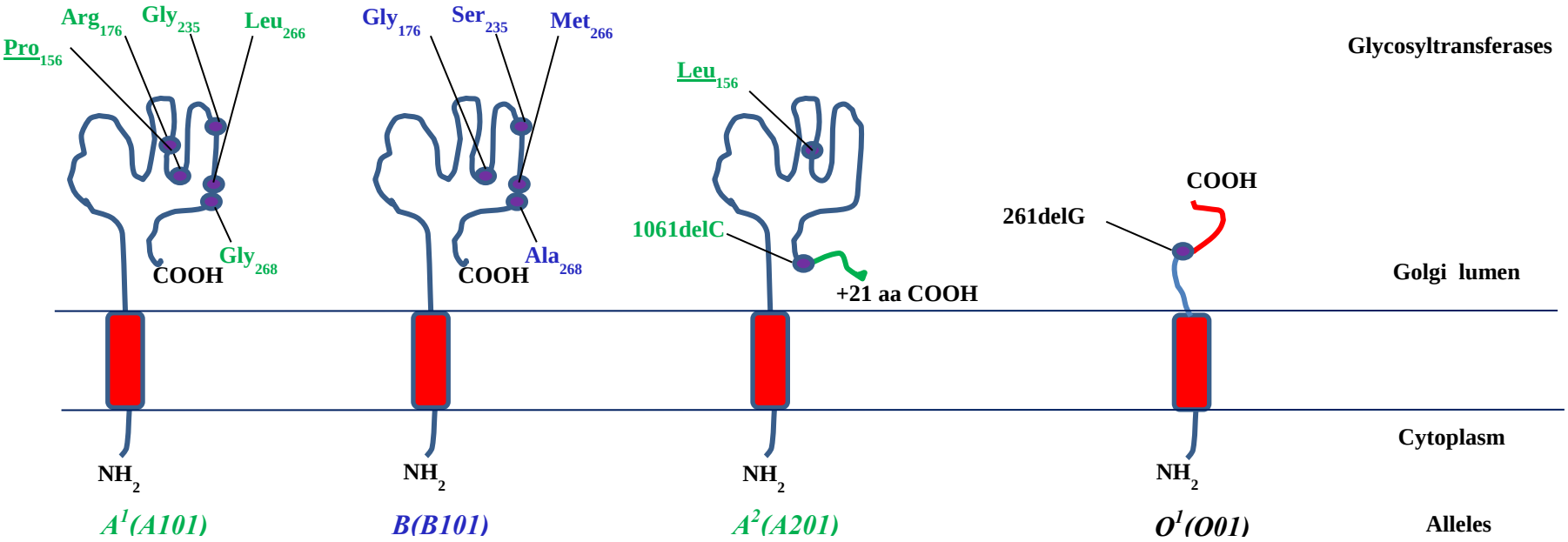


Figure 1.2 Diagramatic illustration of the products (glycotransferases), of four *ABO* alleles, located in Golgi membrane.

The ABO glycotransferases are transmembrane proteins of Golgi apparatus that consist of short *N*-terminal domains (cytoplasmic) and large C-terminal catalytic domains**.** The amino acids (aa) and their locations that differ from those of the A1 transferase are shown. A and B transferase differ by four amino acids at positions 176, 235, 266 and 268 in the C-terminal. A2 transferase differs from A1 by aa change at location 156 (underlined) with a C deletion at 1061 in *A201* allele, leading to 21 aa longer C-terminal. A deletion in exon 6 in *O01* allele leads to a truncated glycotransferase (catalytic domain affected) (adapted from Avent ND 2016 Transfusion and transplantation science in preparation. Oxford University Press).(NOTE the deletion responsible for the *A201* and *O02* alleles occur in the gene not in the glycotransferase)

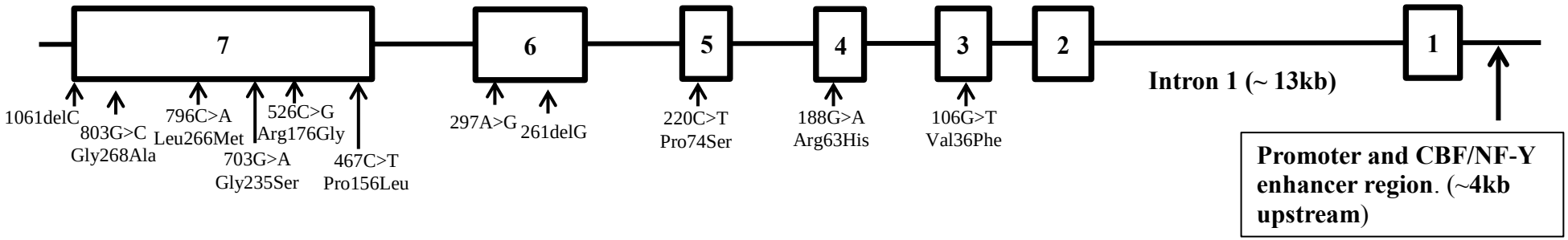


Figure 1.3 The ABO gene (seven exons and six introns).

ABO gene is approximately 20039 bp (According to NCBI GenBank) with 7 exons shown as boxes while the 6 introns are illustrated as a fine line. Crucial polymorphisms, most of which encode for amino acid changes, are illustrated: those that differ between the A^1 (*A101*) and the *B* (*B101*) allele are (Gly268Ala, Leu266Met, Gly235Ser, and Arg176Gly), (1061delC and Pro156Leu) are associated with A^2 (*A102*) allele, the 261delG are with O^1 (*O01*), which is also associated with (*O02*) allele along with Pro74Ser, Arg63His and Val36Phe. The order of the exons is in reverse direction (right to left) due to its illustration as anti-sense direction in the chromosome 9.

1.5.4 *ABO* polymorphisms (*ABO* alleles)

The *ABO* gene is known to be significantly polymorphic, especially in exon 7, giving rise to a considerable number of alleles. Some polymorphisms affect the specificity of the product (glycotransferase) that accounts for the blood groups A, B or O, whereas other mutations might affect the enzyme activity, giving rise to weakened expression of ABO blood group phenotypes (antigens) that are referred to as subgroups. These polymorphisms can occur in various parts of the *ABO* gene with different mechanisms, such as exons, introns and regulatory regions (Chester and Olsson, 2001). There has been a significant growth in the number of the *ABO* alleles due to the continuous emergence of the polymorphisms in the *ABO* alleles. Seventy-two different *ABO* alleles were mentioned in 2001 by Chester and Olsson; this number grew to 215 alleles in 2009 (Storry and Olsson, 2009), and most recently, in August 2016, the Blood Group Antigen Gene Mutation Database, 'dbRBC' reports 381 *ABO* alleles (dbRBC, 2016), some of which might result in different patterns and levels of agglutination in serology, if antigenicity is affected (Yamamoto, 2004).

The common allele of *A (A101*) (often considered the consensus allele) varies from the *B* (*B101*) allele by seven nucleotides in exon 6 (297 A/G) and in exon 7 (526 C/G), 657C/T, 703 G/A, 796 C/A, 803G/C and 930G/A), four of which encode for aa changes Arg176Gly, Gly235Ser, Leu266Met and Gly268Ala in the produced enzymes (Yamamoto et al., 1990a, Daniels, 2005). On the other hand, the most common *O* allele (*O01*) shows an identical sequence to that of *A101* apart from a G deletion in exon 6, which causes a shift in the reading frame altering the aa sequence after aa 88. As a result, a premature stop codon signal is present after aa 117, which encodes for an inactive truncated protein that lacks the catalytic domain of the enzyme (Yamamoto et al., 1990a).

The A antigen is divided into two common subgroups—A_1 and A_2 subgroups—which can be serologically distinguished by the anti-A and anti-A_1 antibodies. A_1 reacts with anti-A and A_1, whereas A_2 reacts with anti-A but not with anti-A_1 (Daniels, 2005). The A_2 phenotype is commonly encoded by the *A201* allele, which differs from *A101* by a missense SNP in exon 7 (467C>T, Pro156Leu) and a deletion in the codon before the translation stop codon in exon 7. Consequently, a disruption in the stop codon occurs, resulting in the production of an enzyme with an extra 21 aa at the C-terminus which are likely responsible for reducing its activity. This leads to possibly lower A antigen expression and a weaker phenotype on the red cells comparing to case with the production of the GTA enzyme in the A_1 phenotype (Yamamoto et al., 1992).

There are considerable *ABO* alleles, some of which result in weak expression of the antigens (subgroups) with various genetic polymorphisms and hybrid genes. The detailed discussion of these alleles, which occur at various frequencies among populations, is not within the scope of this thesis (Reid et al., 2012, dbRBC, 2016, Chester and Olsson, 2001). Another example of the *A1* allele is *A102*, which differs from *A101* by SNP 467C>T, Pro156Leu, and is common among Asians (Reid et al., 2012). With regard to *O* alleles, another common allele is *O02*, which shows a G deletion in exon 6 but differes from *O01* by nine SNPs spread across exons 3-7 (Olsson and Chester, 1996, Reid et al., 2012). Several other weak expression subgroups have been attributed to hybrid alleles (crossover between two different alleles) (Chester and Olsson, 2001). An example of subgroups with weak expression is the A_x phenotype. One possible molecular mechanism for this phenotypes is hybrid alleles of *O02* and *B* or *A*, with exon 7 from the former and the rest from the latter; the contradiction between exon 7 of the O allele (normally encoding for silencing enzyme) without the deleteion in exon 6 leads to an unexpexted phenotype with weak antigen expression (Olsson and Chester, 1998).

Polymorphisms in introns have been reported to be specific for alleles; for example, introns 6 of *A101* and *O101* have been described to be identical, whereas that of *B101* and *O02* differ by 12 and 13 positions, respectively, (Chester and Olsson, 2001). In addition, intron polymorphisms might have an affect on the antigen exprssion. B_m phenotype, which is charaterised by the reduced expression of B antigens on the RBCs, is associated with a deletion in an erythroid-specific regulatory element containing binding sites for the GATA-1 haemopoietic transcription factor. This deletion is located in intron 1 (5.6–6.1 kb from the beginnging of the *ABO* gene) (Sano et al., 2012), which is speculated to cause downregulation of allele (B^m allele) transcription. With regard to the upstream area of the *ABO* gene, it has been pointed out that a minisatellite (sequence repeats) of 43 bp might play a role in transcription reglution since it contains a CBF/NF-Y transcription factor binding site (Kominato et al., 1997). Further, it has been reported that various *ABO* alleles carry different number of repeats as four copies were carried in *A201*, *B101*, *O101* and *O02*, whereas *A101* and *O03* (O^2) have one repeat (Irshaid et al., 1999). Vartiations in these sequences might disrupt the ABO antigens, which has been implicated in the weak expression of the B antigen in two samples that might be as the result of reduced number of the 43-bp repeats (Seltsam et al., 2007).

These various polymorphisms and alleles, along with their distribution, are suggested to be the result of evolution, possibly to tolerate pathogens (Storry and Olsson, 2009, Yamamoto, 2004). The common G deletion in exon 6 is suggested to appear once in human evolution of the more ancient *O02* allele and the *O01* allele arised from exchange between *A101* and *O02* alleles (Roubinet et al., 2004).

1.5.5 *ABO* antibodies (clinical significance)

The antibodies of the ABO system (anti-A and anti-B) are naturally occurring antibodies in the sera of those lacking the antigens (as explained earlier) from the age of six months. They are thought to be the result of exposure to environmental materials, such as bacterial carbohydrates (A- and B-antigen-like), found in bacterial cell walls, that evoke the immune response (Watkins, 2001). This can be seen in the study where children of blood group A were fed parenterally in an attempt to avoid introduction of bacteria, leading to the absence of the expected B antibodies (Storry and Olsson, 2009). In addition, the anti-B activity increased in infants (groups A and O) fed by bacteria (*Escherichia coli* O_{86}), which mainly resembles antigen B activity *in vitro* (Springer and Horton, 1969). Anti-A and anti-B antibodies are mainly IgM, but also may exist as IgG. As a consequence of their characteristics, i.e. natural occurrence and immediate reactivity, A and B antibodies are considered to be the most clinically significant in transfusion. A major mismatch in red cells, for example from donor A to recipient B, could lead to rapid haemolytic transfusion reaction (HTR) (section 1.4.1), which can be fatal (Janatpour et al., 2008). On the other hand, the effect of those antibodies seems to be rare and mild in the case of haemolytic disease of the newborn. Mainly, this occurs in group O mothers of foetuses with other blood groups (A or B) in that the mothers may have low levels of IgG antibodies capable of crossing the placenta and the density of A and B antigens on the foetal RBCs are stated to be low (Daniels, 2013). In addition, since the ABO antigens are expressed in various bodily tissues, the corresponding antibodies may impact the success of solid organ transplantation. Although anti-A and anti-B antibodies might cause hyperacute rejection of solid organs, such as the heart and liver, ABO-mismatched renal transplantation has become common practice. It is accomplished with the treatment of the recipient by reducing the antibody level in the recipient plasma (Daniels, 2013, Storry and Olsson, 2009).

1.6 Duffy blood group system (FY)

The FY blood group system is considered to be a clinically significant blood group since it is reported to be involved in HTR and HDFN cases. The Fy glycoprotein is located on RBCs and other tissues and has been suggested to act as a chemokine and malarial parasites receptor. The FY blood group system was first described in 1950 when the anti-Fy^a to Fy^a antigen was recognised during an investigation of a case of haemolytic transfusion reaction. The patient (surname was Duffy) suffered from rigors and jaundice after receiving three blood units (Cutbush and Mollison, 1950). Subsequently, anti-Fy^b and Fy^b were discovered one year later, reviewd by (Meny, 2010, Westhoff and Reid, 2004). The *FY* gene locus is the first human gene to be assigned to an autosome chromosome (Donahue et al., 1968). Fy^a and Fy^b are inherited in a codominant fashion of *FY*A* and *FY*B* alleles. Other alleles encode weakened antigen expression or abolish the antigen expression (Meny, 2010, Reid et al., 2012).

1.6.1 Duffy glycoprotein and *FY* gene.

Duffy glycoprotein has been suggested to act biologically as a receptor for chemokines, which are signals involved in the interaction between cells required for various processes, such as activation and movement of leucocytes (chemotactic cytokines) (Davenport, 2009). As a result, the Fy glycoprotein was named Duffy Antigen Receptor for Chemokines (DARC). However, the Fy glycoprotein structure has been shown to differ from the conventional (typical) chemokine receptors in terms of lacking the motif Asp-Arg-Tyr/DRY in the second cytoplasmic loop that is required for binding G-protein, leading to lack of signalling ability. Accordingly, the Duffy glycoprotein has been assigned to a new chemokine class namely (Atypical Chemokine Receptor) and named ACKR1, which is also the name of the *FY* gene according to HGNC **(***ACKR1*) (Bachelerie et al., 2014, Bachelerie et al., 2015). The function of the Fy glycoprotein is not entirely known, although some suggest that it acts as a sink or scavenger present on

RBCs for binding the excess plasma proinflammatory chemokines to prevent activation of neutrophils and inflammation (Pruenster and Rot, 2006). This was seen in a study that compared the chemokine levels in the RBCs of Duffy-positive and Duffy-negative volunteers intravenously infused with endotoxin; as compared to the Duffy-negative cohorts, the level of the chemokine was higher in those with positive Duffy expression (2-3 times in plasma) and 20-50 times associated with RBCs (Mayr et al., 2008). Another role of the Duffy glycoprotein is a receptor for the malarial parasites, *P. vivax* and *Plasmodium knowlesi* (Miller et al., 1975*)*. In an experiment, in which 11 patients exposed to *P. vivax* mosquitos were tested, only those with Fy(a-b-) phenotype were found to be resistant (Miller et al., 1976). This has been suggested to occur due to an interaction of the *P. vivax* Duffy-binding protein (PvDbp) with the Duffy antigen on RBC to allow infection (King et al., 2008). As a result, those with the phenotype Fy (a-b-), which are mainly of black African ethnicity (Reid et al., 2012), are most likely resistant to malaria (Miller et al., 1976). Consequently, the *P. vivax* malaria strain is absent in West Africa, where the majority of the population is Fy(a-b-) (Meny, 2010). Fy antigens are reported to be also expressed on the epithelial cells of other organs, including the kidney (Hadley et al., 1994, Chaudhuri et al., 1997).

The Fy glycoprotein (antigen), which is encoded directly by the *FY* gene, is a 336-aa (*N*-glycosylated) protein, spanning the RBC membrane seven times, with extracellular *N*-terminus (glycosylated) and cytoplasmic C-terminus, as shown in Figure 1.4 (Westhoff and Reid, 2004). It is the product of the major and more abundant mRNA form from the *FY* gene.

The *FY* or *ACKR1* gene is located on the long arm of chromosome 1 (1q21-22) (NCBI, 2016b). Two forms of mRNA transcripts have been described, with a single exon in one and two exons in the other (Daniels, 2013) encoding for minor and major forms of the Fy glycoprotein, respectively. Firstly, the *FY* cDNA was cloned and described to be

encoded by one exon gene that produces a 338-aa glycoprotein (Chaudhuri et al., 1995, Iwamoto et al., 1995). Subsequently, it was described that a predominant and more abundant transcript consists of two exons, with the first containing the translation-initiating codon (Met), and separated by a 479-bp intron. This transcript encodes for the major form of the glycoprotein with 336 aa (Iwamoto et al., 1996a). The major and predominant form of the glycoprotein and the mRNA has been preferably used for the numbering of the nucleotides and aa and will be used for this project (Reid et al., 2012, Daniels, 2013, dbRBC, 2016).

1.6.2 Molecular basis of FY blood group system

Fy^a and Fy^b antigens are inherited co-dominantly by *FY*A* and *FY*B* alleles, manifested by three phenotypes, namely Fy (a+b-), Fy (a-b+) and Fy (a+b+). The two alleles differ by a missense SNP at location 125 in exon 2 (G in *FY*A* and A in *FY*B*) that encodes for an aa change on the extracellular domain with Gly42/Fy^a and Asp/Fy^b (Figure 1.4) (Tournamille et al., 1995b, Reid et al., 2012). Other Fy high-frequency antigens are Fy3, Fy5 and Fy6, which are detected by reactions with specific antibodies. Fy3 is found on RBCs of all Duffy phenotypes apart from those with Fy (a-b-) and located on the third extracellular loop of the glycoprotein (Wasniowska et al., 2004) (Figure 1.4). This antigen is defined by anti-Fy3, which is synthesised by immunised individuals with the Fy (a-b-) phenotype; therefore, those with anti-Fy3 antibody should be transfused with Fy (a-b-) red cells (Daniels, 2013). Unlike, $Fy^{a/b}$, Fy3 antigenic determinant is resistant to protease treatment; therefore, anti-Fy3 antibody reacts with enzyme-treated RBCs of Fy^a or Fy^b phenotypes but not those of Fy (a-b-) (Meny, 2010). Fy5 is similar to Fy3 as it is resistant to enzyme treatment but differs by its absence in individuals with Rh_{null} with normal expression of Fy^a , Fy^b and thus Fy3 antigens (Meny, 2010, Daniels, 2013). Anti-Fy6 antibody, which is a murine antibody, recognises a linear epitope located on the *N*-terminal extracellular domain of the Fy glycoprotein between aa 19 and 26

(Waśniowska et al., 1996, Wasniowska et al., 2004, Daniels, 2013). Fy6 is found in all Duffy-positive phenotypes, and unlike Fy3, it is sensitive to enzyme treatment, such as papin (Daniels, 2013).

The expression of Fy antigens has been reported to be weakened or abolished by different polymorphisms and thus alleles (Reid et al., 2012). An example of the weakened Fy antigen expression phenotype is the Fy^x phenotype, Fy ($b+^w$), which is characterised by a reduced expression of Fy^b along with decreased expression of Fy3, Fy5 and Fy6 antigens encoded by the allele on the *FY*B* background (*FY*02M.01*). This allele resembles the backbone of *FY*B* allele but carries a missense SNP 265C>T in exon 2 that leads to an aa change Arg89Cys located in the first cytoplasmic loop of the Fy antigen (Figure 1.4). This might lead to the instability of the glycoprotein, compromising its insertion in the RBC membrane (Tournamille et al., 1998, Yazdanbakhsh et al., 2000). In addition, another aa change, Ala100Thr (due to 298G>A in exon 2), exists in those with Fy^x encoded by *FY*02M.01*, but it is stated that there is no effect on reducing the antigen expression if this change is present alone (Olsson et al., 1998, Gassner et al., 2000). This weakened phenotype might be mistyped by serology as Fy^b negative that might lead to delayed HTR in an immunised recipient (Olsson et al., 1998, Daniels, 2013) Serologically, it has been shown to be impossible to distinguish between the heterozygous *FY*B/X* and *FY*B/B*) (Daniels, 2013). Moreover, the null phenotype Fy (a-b-), which is common among black individuals of African descent (about 70%) has been reported to arise from various polymorphisms (Meny, 2010, Daniels, 2013). One example is the common allele (*FY*02N.01*) among blacks of African origin, which is carried on the *FY*B* background (identical to *FY*B*) but with SNP T>C 67bp upstream (-67) of the major translation start codon, which is in the GATA promoter region, disrupting the binding site of the erythroid transcription factor (GATA-1) (Tournamille et al., 1995). Consequently, since this erythroid promotor

controls the antigen expression on RBCs, those with this allele lack the Fy antigen expression on RBCs but not other cells, such as endothelial cells in the spleen (Peiper et al., 1995, Iwamoto et al., 1996b). Accordingly, those with this phenotype are believed to possess only anti-Fy^a antibodies but not anti-Fy^b antibodies (Westhoff and Reid, 2004). Although it has been associated with *FY*B*, this GATA mutation has subsequently been seen on the *FY*A* background (*FY*01N.01*) in Papua New Guinea (Zimmerman et al., 1999). The null phenotype is reported to be rare in ethnic groups other than those of black origins (Daniels, 2013, Reid et al., 2012). Those arising from various polymorphism mechanisms such as nonsense SNP, leading to stop codon or nucleotide deletion. For example, the allele *FY*01N.02* carries a 14-bp deletion in exon 2 in *FY*A*, resulting in a reading frame shift and stop codon in Caucasians (Mallinson et al., 1995). In addition, a nonsense SNP leading to premature stop codon has been reported to occur in both *FY*A* and *FY*B* backgrounds of Caucasians in exon 2. For *FY*A*: 408G>A and 287G>A, to account for the silencing alleles *FY*01N.03* and *FY*01N.04*, respectively, while 407G>A accounts for *FY*02N.02* (*FY*B* background). These alleles encode for truncated protein, that unlike that of black Africans, leads to loss of expression in all cells, including RBCs (Rios et al., 2000). According to BGMUT database, there are 16 variant *FY* alleles with various polymorphisms that might affect the expression, and the number is expected to continually increase due to the discovery of novel SNPs and alleles not yet included; for example, two novel silencing *FY*B* alleles were described in 2014 to result from 2 nucleotides deletion and a missense SNP 895G>A leading to Ala299Thr, respectively (Westoff et al., 2014, dbRBC, 2016).

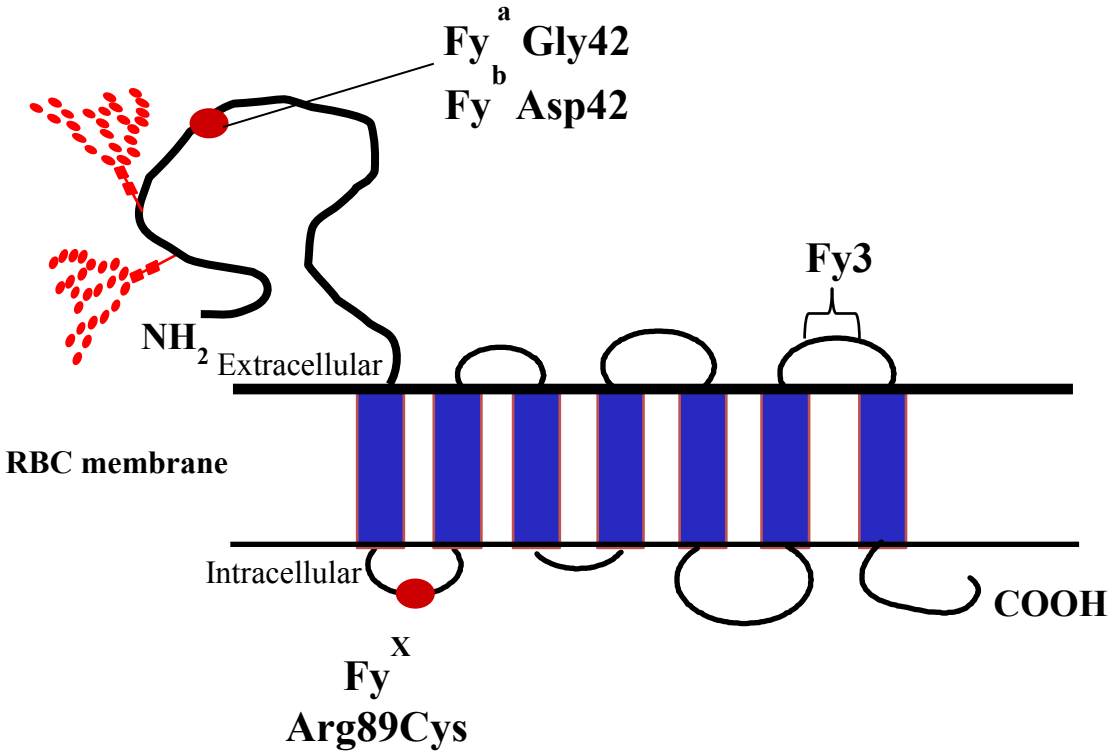


Figure 1.4. The structure of the Duffy protein (7-transmembrane domain).

The Fy glycoprotein spans the RBC membrane seven times, with extracellular *N*-terminus (glycosylated) and cytoplasmic C-terminus. The aa changes for the Fy^a/Fy^b and Fy^x are shown. In addition, Fy 3 location, which is detected by antibody, is displayed (adapted from Avent ND 2016, Transfusion and Transplantation Science, Oxford University Press).

1.6.3 FY antibodies (clinical significance)

FY antibodies exist as a consequence of immunisation via exposure to the corresponding antigens acquired through blood transfusion or, rarely, pregnancy (Meny, 2010, Daniels, 2013). They are mostly IgG, IgG1 and, rarely, IgM (Westhoff and Reid, 2004), and they could lead to immediate and delayed HTRs (Poole and Daniels, 2007) that range from mild to fatal, especially anti-Fy^a and $–Fy^b$ antibodies (Daniels, 2013). Therefore, it is suggested that red cells with the negative Fy antigen should be given to those with pre-existent antibodies.

With respect to HDFN, anti-Fy^a and anti-Fy^b antibodies usually lead to mild reactions, although it can be severe, as seen in the case of three incidences out of 68 cases possessing anti-Fy^a requiring intrauterine transfusion due to anaemia in the foetus (Goodrick et al., 1997).

1.7 Kidd blood group system (JK)

The JK blood group system is considered to be among one of the most clinically significant groups after ABO and Rh systems; the antibody pertinent to this system might lead to transfusion reactions or HDFN. The discovery of the Kidd blood group system was made in 1951 when HDFN was investigated in a patient (Mrs Kidd) leading to the detection of a new antibody (anti-Jk^a). The anti-Jk^b antibody was described two years later in an individual with a transfusion reaction (Plaut et al., 1953, Westhoff and Reid, 2004). Subsequently, the anti-Jk3 antibody was detected in individuals with phenotype Jk (a-b-) and was found to react with RBCs with both Jk^a and Jk^b antigens (Irshaid et al., 2000). Accordingly, three Jk antigens are recognised by ISBT for the JK system (ISBT, 2016, Reid et al., 2012) (Jk^a, Jk^b (polymorphic) and Jk3).

1.7.1 JK glycoprotein and the *JK* gene

The JK glycoprotein is suggested to act as a urea transporter, as shown in a urea lysis test. It has been reported that RBCs with the phenotype Jk (a-b-) are resistant to lysis, while those with the common Jk phenotype are rapidly lysed in 2 M urea (Heaton and McLoughlin, 1982, Daniels, 2013). Kidd glycoproteins are distributed on red cells and other tissues, such as kidney, brain and heart tissues (Olivès et al., 1996). The glycoprotein is an integral protein spanning the membrane ten times with both the *N*- and C-terminal being intracellular and *N*-glycosylation being present at Asn211, which is on the third extracellular loop (Lucien et al., 2002b) (Figure 1.5).

The *JK* gene (*SLC14A1*) is located on chromosome 18 at 18q12.3 (NCBI, 2016b) and consists of 11 exons distributed over about 30 kb. The translation start codon is located in exon 4; therefore, exons 4–11 account for the production of the mature protein, which consists of 389 aa (Cartron et al., 1998, Lucien et al., 1998). In addition, -837 to -336 upstream contains the erythroid-specific GATA-1 transcription factor binding site (Lucien et al., 1998).

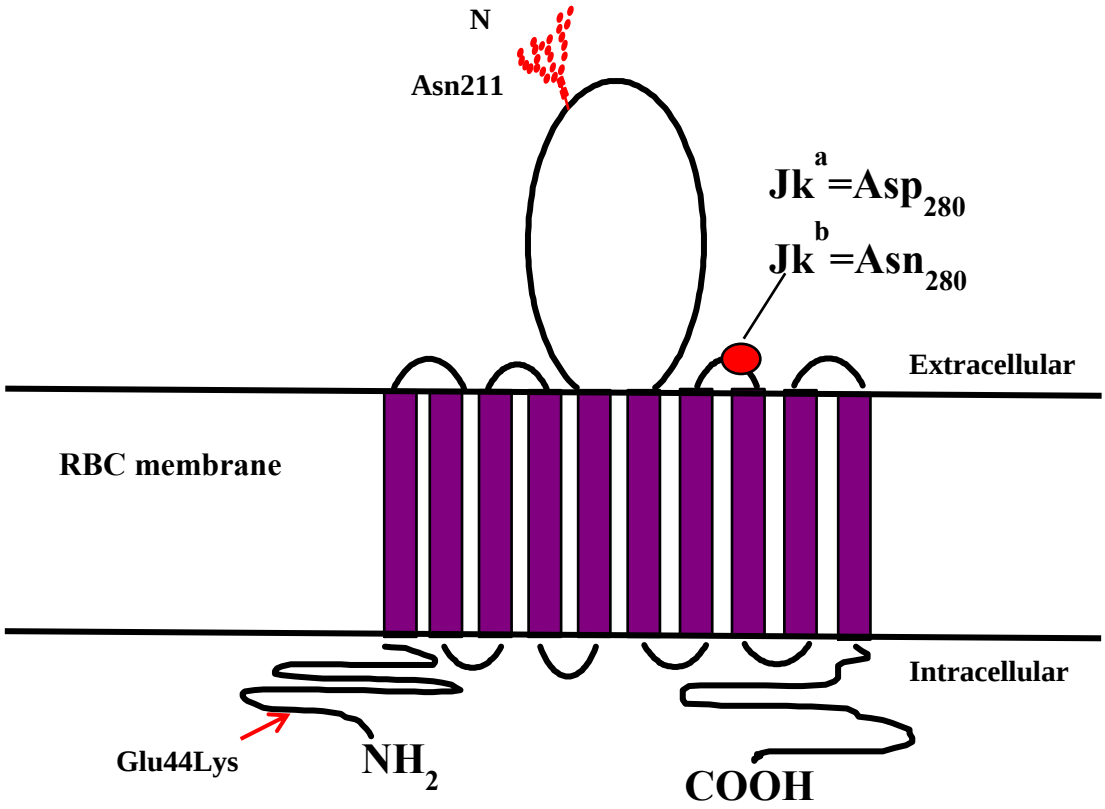


Figure 1.5 The Kidd glycoprotein, showing 10 membrane spanning domains.

The polymorphism of Jk^a/Jk^b, is located on the fourth loop. Glu44Lys, is associated with weak Jk^a expression, which is also shown. The N-glycan on the extracellular loop is also displayed. (Adapted from Avent ND 2016, Transfusion and Transplantation Science, Oxford University Press).

1.7.2 Molecular basis of JK blood group system

Antigens Jk^a and Jk^b are inherited via a co-dominant mechanism by the *JK*A* and *JK*B* alleles, respectively (Daniels, 2013). *JK*A* and *JK*B* alleles differ by a missense SNP (838G>A) in exon 9 that encodes for the aa change Aps280Asn that is located on the fourth extracellular loop of the Jk glycoprotein. It has also been pointed out to differ by a silenct SNP in exon 7 (588A in *JK*A* and 588G in *JK*B*) (Olivès et al., 1997, Irshaid et al., 2000). The Jk3 antigen is found on Jk^a and Jk^b RBCs, although the responsible specific aa is unknown (Westhoff and Reid, 2004) (Figure 1.5).

The inheritance for the Jk null phenotype has been suggested to be due to homozygous or heterozygous inheritance of recessive silencing alleles with inactivating mutations and is rare in some ethnic populations but abundant in Polynesians (Reid et al., 2012). Those with this phenotype produce anti-Jk3 antibodies upon immunisation (Daniels, 2013) and might require Jk3-negative units (Irshaid et al., 2000). Several polymorphism mechanisms have been reported to account for silencing alleles which could be on the background of *JK*A* or *JK*B*, such as SNPs (missense and nonsense) in exons, splice sites and exon deletion (Reid et al., 2012, Daniels, 2013). Two different SNPs among two populations were described in a study by Irashid et al (2000): one occurs in the invariant acceptor splice site on intron 5 (G>A) that led to skipping to exon 6 in Polynesians, and the other was a missense SNP (871T>C, Ser291Pro) observed in Finns; both were expected to abolish the expression by affecting glycosylation (Irshaid et al., 2000). The abolished antigen expression has been shown to be also caused by exon deletion, for instance, the deletion of exons 4 and 5 has been reported to account for the null Jk phenotype as the translation initiation codon in exon 4 is absent, affecting the production of the Jk glycoprotein (Irshaid et al., 2002, Lucien et al., 2002a). On the other hand, it has been suggested that an inheritance of a dominant gene *In*(*JK*), which appears to be unrelated to the *JK* gene might account for a silent phenotype although the

molecular basis is yet unknown (Daniels, 2013).

Weak expression of Jk antigens (Jk^a or Jk^b) has been reported and might be due to polymorphisms, particularly SNPs, in various parts of the coding areas (exons) within diverse alleles (Reid et al., 2012). One example is the allele *JK*01W.1* described by Wester et al (2011), which occurs on a *JK*A* background and results in reduced expression of mainly the Jk^a antigen. This allele holds a missense SNP (130G>A) that leads to aa change Glu44Lys that is suggested to locate in the cytoplasmic *N*-terminal domain (Figure 1.5) along with *JK*B*-like silent SNPs 588G and -46G in intron 9. Homozygous or heterozygous inheritance of that allele has been shown to react weakly with anti-Jk^a antibodies, that might lead to false-negative interpretations of Jk^a and possible HTR (Wester et al., 2011). Different other polymorphisms have also emerged, leading to 36 *JK* alleles reported so far in the BGMUT database (dbRBC, 2016).

1.7.3 JK antibodies (clinical significance)

Both antibodies Jk^a and Jk^b, which are mainly IgG or mixed with IgM, have been reported to cause HTRs that can be severe (Hussain et al., 2007, Daniels, 2013). They are more commonly associated with delayed HTRs due to their tendency to drop to undetectable levels in the plasma (Schonewille et al., 2006). It is estimated that around one-third of delayed HTRs (DHTRs) were due to anti-Jk^a (Pineda et al., 1999). In addition, anti-Jk3 antibody has been reported to cause severe immediate and delayed HTR (Marshall et al., 1999, Daniels, 2013). As a result, matching blood units (antigen-negative) for those with Kidd antibodies is preferable. On the other hand, with respect to HDFN, Kidd antibodies are rarely associated with severe cases, although a few severe cases have been reported (Daniels, 2013, Ferrando et al., 2008).

1.8 Blood group genotyping

1.8.1 Serology and genotyping

The determination of the individual blood group type is of great clinical importance for the safety of blood transfusions and organ transplantations. Identifying the blood group type has been routinely accomplished through simple serology utilising the agglutination process. The RBC agglutination is inhibited due to the negative charge on their surface owing to the sialic acid residues. Specific antibodies react with antigens on the RBCs to bridge the gap between adjacent RBCs causing agglutination (Pamphilon and Scott, 2007). The direct antiglobulin test (DAT) is used to investigate whether RBCs are sensitised by antibodies, for example in autoimmune haemolytic anaemia, in which antihuman globulin is added to patient's sensitised RBCs causing agglutination. On the other hand, indirect antiglobulin test (IAT) can be used for matching blood units, for instance; the antibodies in the recipient plasma are tested against the donor's RBCs, when they are added together, followed by the addition of antihuman globulin to check for agglutination (Coombs and Roberts, 1959).

Despite the fact that the serological approach has been found to be reliable, cheap, simple and sensitive in defining the blood group status in individuals and been the gold standard approach for the transfusion practices, serological typing holds limitations that can be overcome by DNA-based analysis. Examples of such circumstances include determination of the blood group of multi-transfused patients, which can possibly carry mixed blood antigenicity from donor's RBCs, leading to inaccurate serological typing (Reid et al., 2000). In addition, serological typing is difficult in the case of limited availability of manufactured antisera for a number of infrequent blood groups (like the case of Dombrock system) (Baumgarten et al., 2006). Other drawbacks are difficulty in addressing discrepancies and in interpretation of unexpected weak antigen expression (phenotype) due to allelic variation, especially rare, at the molecular level (McBean et

al., 2014). Moreover, on a large scale, serological typing can be labour intensive and expensive in terms of the need of a large number of reagents to include a wide range of antigens. As a result, the high-throughput typing for donors using serology is unfeasible, which results in a low number of inventories for antigen-negative blood (matched blood especially for multi-transfused individuals) (Reid, 2009). Therefore, genotyping, particularly high throughput, has been considered for typing the blood group antigens of patients and donors in order to provide extensively matched blood units. Accordingly, this might lead to a fully typed donor database, preventing further transfusion reactions (Avent, 2009, Veldhuisen et al., 2009); thus, genotyping would be superior to serology if performed in high throughput and with high accuracy. In addition, the phenotype of individuals with IgG-coated RBCs is difficult to be assigned by serology (Reid, 2009). The limitations of serology have been suggested to be solved by blood group genotyping (BGG), which is being used for such circumstances, but generally to support serology (Avent, 2009).

The molecular background, the gene sequence and polymorphisms affecting antigen expression of the majority of the human blood groups, particularly the clinically significant ones, has been defined. This knowledge regarding these topics has accelerated since the cloning of the major blood group genes RH (Avent et al., 1990) and ABO (Yamamoto et al., 1990a). A large amount of molecular information, including various genetic mechanisms, such as SNPs (see section 1.3) account for a great number of emerging alleles, some of which affect the expression of the antigens. The number of these polymorphisms and alleles is increasing; the number of alleles identified previously was 1568, which has increased to 1779 at present and is expected to grow further (McBean et al., 2014, dbRBC, 2016). As a result, these vast variations in antigen expression pose challenges to comprehensive analysis of blood groups via serology. For example, a significant number (over 50) of monoclonal anti-D is

necessary to cover the D epitopes in order to differentiate among partial D variants phenotypes (Avent, 2009). As in the case of RhD and other blood groups, those variants, such as those epitope missing in the DVI variant, could elicit the immune response if missed in the testing of red cells (Avent, 2009). Therefore, the use of DNA-based analysis (genotyping) has become more feasible to predict the individual blood group phenotype and secure more comprehensive information about individual blood groups than serology alone (Reid and Denomme, 2011, Avent, 2007). Consequently, considerable efforts have been placed to develop new systems and platforms to enhance the ability of utilising the DNA from donors and patients in performing large-scale, high-throughput BGG, which would provide more comprehensive and extended information of the phenotype than that provided by serology (Avent, 2009).

1.8.2 Applications of BGG

Many advantages can be achieved by using BGG over conventional serology upon the knowledge of the molecular information of the blood group systems. It has been suggested that BGG can be utilised instead of serology if the RBCs are not available to assess the phenotype and if molecular data would provide better knowledge than serology in a more cost- and labour-effective and efficient manner (Daniels, 2013).

A few clinical applications of the BGG have been defined, one of which is to reduce and manage the cases of HDFN. The foetal blood group is determined via genotyping the cell-free DNA (cffDNA) found in the maternal plasma. This procedure prevents the invasive procedure of obtaining the sample, eliminates the unnecessary prophylactic administration of anti-D antibodies antenatally to those carrying a foetus negative for Rh or other clinical blood groups and, most importantly, reduces the risk of HDFN (Daniels et al., 2009).

A number of patients, such as those with sickle cell disease, require regular blood transfusion in order to survive. Nevertheless, receiving multiple blood donations is problematic since it may cause alloimmunisation against various blood antigens or high-frequency antigens, resulting in difficulties in providing matched blood units. Extended blood group typing would be difficult by serology, especially due the possibility of the presence of the transfused RBCs in the circulation. Therefore, genotyping, especially high-throughput, is suggested to allow extended blood group typing of a large number of donors that will increase the inventory for compatible blood for such patients and prevent further immunisations (Reid et al., 2000, Avent, 2009, Wilkinson et al., 2012). Otherwise, multi-transfused patients may develop DHTRs because the antibodies corresponding to blood groups are difficult to detect by cross-matching serologically. Some examples of such antibodies are those of the JK and FY system (Avent, 2009).

The extended genotyping and matching might also be beneficial in organ transplantation, particularly in the case of antibodies against antigens known to be expressed in other cells (histo-blood groups); for example, it has been reported that a mismatched FY renal transplant resulted in increased signs of chronic lesions (Lerut et al., 2007). Other applications of BGG are determining the frequency of blood group polymorphisms in a population, screening blood donors for rare phenotypes, determining the blood group in autoimmune haemolytic anaemia and addressing difficult serological cases, particularly in blood group reference laboratories, by providing more accurate prediction of the phenotype (Anstee, 2009).

1.8.3 High throughput

Considering the many applications in which genotyping is superior to serology, it may be inferred that a genotyping approach with high-throughput capacity is necessary in order to address the growing demand for routine BGG (Stabentheiner et al., 2011). The routine extended typing (to include more blood groups other than ABO and RHD) for a

large number of donors in order to establish a database of matching blood units, especially for immunised individuals, is believed to be difficult (large number of anti-sera required, non-availability of anti-sera or weakly expressed antigens that might not be detected) and costly by serology and suggested to be replaced by high-throughput genotyping approach in the future due to it importance in such cases (Perreault et al., 2009, Jungbauer et al., 2012). The high-throughput approach for extended genotyping at a large scale has been shown to be feasible due to its rapid, reliable, cost-efficient and accurate results, unlike the case with serology (Jungbauer et al., 2012). The high-throughput methods continue to be developed and improved in terms of accuracy, cost effectiveness, and ease of processing and analysis and might provide more informative prediction of the phenotype from the genotype (Avent, 2009).

1.8.4 Genotyping technology and methodology

With the knowledge of the molecular basis of blood group systems gained since the cloning of blood group genes, BGG has become feasible. Several methods have been developed for BGG, mainly including PCR-based assays. Examples of these approaches are polymerase chain reaction (PCR) coupled with restriction fragment length polymorphism (RFLP-PCR), to analyse amplified relevant sequences of blood group genes and SNPs of interest (Fukumori et al., 1995); sequence specific primer-PCR (SSP-PCR) and real time PCR (RT-PCR) (Polin et al., 2008). Although these methods have been reported to be accurate and reliable for genotyping, they are not considered high-throughput platforms due to their limited capacity and output that hinders large-scale genotyping (McBean et al., 2014, Veldhuisen et al., 2009). Moreover, some of these methods require post-PCR gel based analysis that might lead to contamination in addition to the difficulties in automation (Veldhuisen et al., 2009). An example for commercially available blood group genotyping platform is RBC-FluoGene (from Inno-

train). This platform is based on the approach of SSP-PCR with the TaqMan® probes for the fluorescence detection of various blood group antigens in automation manners and without the need for post run analysis with agarose gel. Different kits (plates) that cover predefined blood group antigens, such as some of ABO antigens in one kit whereas another kit covers only Jk^a, Jk^b, Fy^a, Fy^b, $Fy^{a-, b-}$ and Fy^x are provided (Inno-train, 2017). Although this platform is fast (90 minutes) and automated, the fact of the reliant on predefined mutations and the low throughput pose limitations to meet the demand for routine BGG.

It has been suggested that a DNA micro-array-based approach might overcome these limitations with the advances in high-throughput genotyping of blood group antigens (Bugert et al., 2005, Beiboer et al., 2005). The DNA microarray technology is characterised by a high multiplexing capability using solid surface that can be glass, chip or bead arrays. These DNA arrays carry numerous probes that are specific to various blood group alleles. During the microarray process, an allele-specific hybridisation occurs between the labelled, amplified PCR products and the corresponding probes, which is measured in order to predict the phenotype (McBean et al., 2014).

There are several commercially available platforms that utilise the microarray technology such as BloodChip reference and human erythrocyte antigen (HEA) BeadChip.

The BloodChip (from Progenika) was developed between 2003 and 2006 by the Bloodgen project, funded from a European Framework V grant, in order to establish an array-based platform that covers a great number of blood group polymorphisms (Avent et al., 2007). The PCR-amplified DNA products are fluorescently labelled and then hybridised to probes that are specific to known blood-group SNPs and alleles attached

on a glass array. Hybridisation to the allelic probes is then detected by a laser scanner, with software that then analyses the genotype and predicts the phenotype. It is suggested that the BloodChip reference version v4.1 can detect up to 33 clinically significant antigens from ten different blood group systems (Avent et al., 2009).

The HEA BeadChip, developed by BioArray Solutions, principle is similar to that of the abovementioned platform but has probes attached to a colour-coded bead assembled into arrays on chips, then fluorescent signals from hybridisation are detected to predict phenotypes. It is suggested that the BeadChip platform can genotype SNPs within 18 alleles associated with 24 blood-group antigens from 10 blood group systems, such as FY and JK. In addition, this platform is suggested to be capable of genotyping up to 96 samples in four hours (Hashmi et al., 2005, Hashmi et al., 2007). Another platform, MassARRAY® System, was introduced by (Agena Bioscience) that can provide high-throughput BGG. Two panels, namely, Hemo ID™ Blood Group Genotyping and Donor Quick Screen (DQS) panels can be used with this system to enable genotyping of up to 101 antigens of 16 blood group systems for up to 3000 samples. These panles cover predefined polymorphisms. Briefly, the process involves multiplex PCR followed by loading the samples to a SpectroCHIP® Array before detection by the mass spectrometry (Agena Bioscience, 2017a, Agena Bioscience, 2017b).

Although these platforms have revolutionised the genotyping approach with their high-throughput ability, they are only capable of detecting known SNPs and alleles included in the system array. With the emergence of novel or rare SNPs and combinations of other mechanisms of polymorphisms in alleles of clinical significance, which can alter antigen expression of blood group systems, current microarray platforms fail to detect those not included in the system array. For example, microarray platforms might provide false prediction of normal expression phenotype for null or weak phenotypes due to presence of silencing SNPs not included in the system. Consequently, the

predicted phenotype from an apparent normal allele might contradict with the serological expression. As a result, constant update to the array-based platforms for new alleles, especially if they are clinically significant, is required; this would be costly and labour intensive (McBean et al., 2014, Tilley and Grimsley, 2014, Avent et al., 2015).

Consequently, sequencing-based typing (SBT) has been applied to overcome these constraints since it analyses the entire gene, nucleotide by nucleotide, to provide the order of the four nucleotides: adenine (A), thymine (T), cytosine (C) and guanine (G). As a result, with the molecular knowledge of blood groups, novel alleles are likely to be identified. In fact, Sanger sequencing has increasingly been used to solve discrepant results of serological phenotype and those predicted by genotype approaches (mentioned above) resulting from rare or null polymorphisms or complex genetic variations (an example is shown in JK chapter, section 4.4.4). Although Sanger sequencing might be able to detect novel alleles, it is not considered to be high throughput; this could be a significant limitation in light of the massive demand for genotyping of donors and emerging alleles. As a consequence, sequencing-based typing with the ability of high-throughput sequencing is suggested to be the great candidate to address the importance of genotyping donors and patients (Avent et al., 2015).

DNA sequencing has been based on Sanger's chain-terminating dideoxynucleotide sequencing with capillary electrophoresis, reviwed by (Shendure and Ji, 2008). The original Sanger sequencing principle was based on chain-terminating dideoxynucleotide triphosphates (ddNTPs) which lack the 3′ hydroxyl group required for the phosphodiester bond between two nucleotides. In four separate reactions, single-strand DNA template, polymerase and normal dNTPs initiate growth (sequence by synthesis) of the DNA chain until the incorporation of the terminating ddNTPs that stop the extension. These are loaded onto four adjacent lanes on a polyacrylamide electrophoresis gel, and the sequence is visualised and analysed according to size

separation and nucleotide type, with the shortest sequence fragments appearing on the bottom and longest ones on the top (Sanger et al., 1977, Mardis, 2013). Then, fluorescent dye-labelled ddNTPs (distinct for each ddNTP, to allow simultaneous addition of all ddNTPs in one reaction) were introduced (Prober et al., 1987) along with the capillary electrophoresis that enabled automation (Wong, 2013). Each ddNTP has a unique fluorescent dye that stops the reaction of the primer extension when integrated to the PCR amplicon strand. This results in products of different sizes, depending on the order of the corresponding nucleotides that subsequently flow, and the products are separated, according to size, in capillary electrophoresis. They are detected, based on their wavelength, by laser technology which produces a visible electropherogram that represents the PCR sequence (Wong, 2013). The number of capillaries has increased to 96 or 384, allowing simultaneous sequencing to increase the throughput. It has been suggested that Sanger sequencing can sequence up to 1000 bp per run per capillary at a high level of accuracy (99.999%) with a cost of $0.50 per kilobase (Shendure and Ji, 2008). Therefore, one run of 96 capillaries would generate 96 kb, which would have to be repeated 31,250 times to sequence 1X coverage of the human genome. This is labour intensive, expensive and time consuming. This can be seen in the Human Genome Project (HGP) that started in 1990, in which automated Sanger-based technology was used with capillary electrophoresis (Liu et al., 2012). A draft of the human genome (90%) was published with a cost of about $300 million and was produced over a 15-month period. A full human genome (about 3 billion bases) was then completed in 2003 and is estimated to have cost ~$150 million (Consortium, 2004). Therefore, it is estimated that since the beginning, the HGP cost about $3 billion and required 13 years to complete (NHGRI, 2016). Thus, although Sanger sequencing is accurate (99.999%), the limited capacity is insufficient, expensive and time consuming to address the demand of large-scale genotyping. Consequently, it would be wise to develop a new

sequencing technology that can overcome these limitations (Metzker, 2010). The high-throughput ability has been suggested to provide time- and cost-effective output. Next-generation sequencing technology (NGS) can overcome the limitations of Sanger technology, thereby reducing the costs of human genome sequencing and to eventually as little as 1000 dollars, as per a long-term goal of National Human Genome Research Institute (NHGRI), by utilising the technology's massive ability of producing a significant amount of data per run rapidly. These benefits are mainly attributed to the mechanism of massively parallel sequencing per run (Wong, 2013, Voelkerding et al., 2009). In fact, it has been reported that the human genome sequencing cost decreased significantly since sequencing centres have switched to NGS technology from automated Sanger (Wetterstrand, 2016). In 2008, a human genome was sequenced by massively parallel sequencing (NGS platform) Genome sequencer FLX instrument 454 at a cost of less than $1 million within 2 months (Wheeler et al., 2008). This shows the great advancement of NGS capabilities, which is still improving, compared to that Sanger used in 2003 (mentioned earlier). Owing to its comprehensive and large-scale capability, the technique will be cost effective, not only in terms of a single run but also in terms of eliminating the costs of transfusion complications and other further analysis (Avent, 2009).

1.8.5 Next-generation sequencing

Following the human genome project (HGP), the high-throughput sequencing approach was considered to produce large-scale data rapidly at low costs. Accordingly, various companies have participated in the race to develop NGS platforms due to the importance of the technology and potential for application in many medical fields. Examples of the resultant platforms are 454 by Roche, which was the first commercially available NGS platform (Margulies et al., 2005), SOLiD (Sequencing by Oligo Ligation Detection), Illumina (Solexa technology) and Ion Torrent Personal Genome Machine

PGM™ (Life Technologies), all of which are based on the massively parallel sequencing principle (which is also used to describe NGS). NGS platforms are diverse in their chemistry of sequencing and detection (Liu et al., 2012). The Roche product, 454 for example, relies on sequencing-by-synthesis, particularly pyrosequencing, which is the detection of pyrophosphate released during nucleotide incorporation. On the other hand, Ion PGM™ detects the pH that changes as a result of the production of hydrogen ions when nucleotides are incorporated (Rothberg et al., 2011). In addition, platforms differ by their capacity, speed and costs, upon which platforms are selected in laboratories with the application size needed, such as whole genome or targeted genes. For instance, the output of SOLiD is suggested to be 120 Gb, whereas that from llumina HiSeq 2000 is 600 Gb (Quail et al., 2012, Liu et al., 2012). Examples of different NGS platforms with their sequencing chemistry are shown in Table 1.3.

Although NGS platforms may differ in their sequencing chemistry, the majority of these innovations share a similar workflow. Following the extraction of the genomic material (here was DNA), the sequencing library is constructed. The targeted region of the DNA, defined by PCR (amplicon), is fragmented by enzymatic shearing or physical sonication. Then, the fragmented DNA is ligated to adapters (both ends), which are unique sequences for each platform such that they are complementary to the sequence on the surface they attach to for clonal amplification. The subsequent step is template preparation, which involves clonal amplification. The sequencing library (ligated) is immobilised on beads or solid surface, depending on the platform for commencing the clonal amplification. Depending on the platform, the clonal amplification can be accomplished by emulsion PCR or solid phase amplification (bridge amplification PCR). As a result, each DNA fragment is amplified to increase the intensity of the signal generated from the sequencing for sufficient and sensitive detection (Metzker, 2010, Shendure and Ji, 2008, Mardis, 2013).

The development in sequencing technology has led to the emergence of the single-molecule long-read sequencing approach, which is referred to as third-generation sequencing. PacBio and Oxford Nanopore Technology (ONT) have been introduced. This technology has several advantages; for example, no clonal PCR amplification is needed, which might reduce the duration for preparation. In addition, single-molecule DNA is sequenced, which is useful in assigning the location of the mutation cis/trans effect of polymorphisms to individual blood group alleles because PacBio and ONT allow sequencing of up to 50 kb and 200 kb, respectively, which is greater than the size of most human blood group genes (Avent et al., 2015, Goodwin et al., 2016). The Ion Torrent (PGM^{TM}) was the platform of choice in the current project since it was available in the research lab in Plymouth University. Since the NGS approach has been used in various applications and fields, it has been utilised specifically for BGG, which is the relevant research to this project.

Table 1.3 Examples of different NGS platforms with their sequencing chemistries. (Adapted from (Hodkinson and Grice, 2015)

NGS Platform	Clonal amplification	Sequencing chemistry	Highest average read length
454	Emulsion PCR	Pyrosequencing (sequencing by synthesis)	700 bp
SOLiD	Emulsion PCR	Oligonucleotide 8-mer chained ligation (sequencing by ligation)	75 bp
Ion Torrent	Emulsion PCR	Proton detection (sequencing by synthesis)	400 bp
Illumina	Bridge amplification	Reverse dye terminator (sequencing by synthesis)	300 bp

1.8.6 Ion Torrent Personal Genome Machine™ (Ion PGM™)

Ion PGM™ was developed by Ion Torrent and then purchased by Life Technologies™ and is currently owned by Thermo Fisher Scientific. It was introduced in the market in 2010 by Ion Torrent as a feasible, affordable, reliable and fast (running time approximately from 2 to 7 hours depending on the capacity needed) sequencing platform that can be used for smaller laboratories (Hodkinson and Grice, 2015, ThermoFisher, 2015). It is based on semiconductor technology and pH detection chemistry of sequencing by synthesis, unlike other NGS platforms that utilise light or camera detection. During the sequencing, each nucleotide is incorporated into the nascent strand of DNA by a polymerase leading to hydrogen ion (H^+) release. As a result, the pH of the solution surrounding the bead changes (0.02 pH unit per base incorporation), which is detected by a sensor (works as a solid-state pH meter) underneath converting the change into a voltage signal that is digitalised as base calling information by off-chip electronics within only 4 seconds. The four nucleotides are flooded in a stepwise fashion into wells containing the templated beads. In each nucleotide flow, if a nucleotide is incorporated, a signal is reordered; however, if more than one, for instance two nucleotides are incorporated, the voltage is doubled. A wash step between each nucleotide flow is conducted to remove the remaining nucleotides from the well (Rothberg et al., 2011, Mardis, 2013, ThermoFisher, 2016). Figure 1.6 illustrates the sequencing reaction and detection principle of Ion PGM™.

The workflow consists of library preparation, template preparation and sequencing. The DNA library is produced by targeting the genomic region by PCR (in this case, long-range PCR), followed by fragmentation by enzymes, which can also be alternatively be achieved physically (by sonication). Then, these fragments are ligated with adapters that can be recognised by complementary sequences on the beads. During their preparation, templates on the beads undergo clonal amplification by emulsion PCR—a process in

which each bead is confined in a droplet surrounded by oil, resulting in millions of uniform DNA fragments in order to increase the signal produced. Then, the oil is separated from the beads (emulsion breaking), disassociating the complement strand and leaving single-strand DNA. Finally, with the DNA polymerase and sequencing primer, the enriched beads (beads with insufficient DNA are removed) are loaded into the chip wells and sequenced as explained earlier (Mardis, 2013, Goodwin et al., 2016) (Figure 1.7). This sequencing process occurs in each well of the chip, which is designed to use the complementary metal oxide semiconductor (CMOS) process and contains millions of 3.5-mm-diameter wells (depending on the chip size, 314, 316 and 318), with a pH sensor underneath each well. Thus, simultaneous detection of massively parallel independent sequencing reactions (massively parallel sequencing principle) is allowed, resulting in high-speed and throughput and low-cost sequencing (Rothberg et al., 2011).

The sequencing capacity and output depends on the chip used and therefore, on the well - the number of wells as well as the read length (the longer the read length, the greater is the output by the chip). For example, the Ion chips 314™, 316 ™ and 318 ™ for the Ion PGM ™ consist of 1.2, 6.3, and 11.3 million wells that would allow an output of 20 Mb, 200 Mb and up to 1 GB for read length of 200 bp, respectively (Merriman et al., 2012). The capacity has been increased by improving the chip's capacity with a second version of Ion Chips that allows read length of up to 400 bp with an output of up to 2 Gb (ThermoFisher, 2015) (Table 1.4). Based on the output, the Ion PGM ™ is considered beneficial for small laboratories with smaller projects such as targeted DNA and RNA sequencing, gene sequencing and genotyping (ThermoFisher, 2015). On the other hand, a new platform (Ion Proton™) by Life Technologies has much higher capability than Ion PGM ™ and is suitable for large projects. A whole human genome and up to two human exomes (exons of the human genome) can be sequenced within two to four hours. The Ion Proton™ II chip consists of 660 million wells, allowing a significantly

higher output of up to 10 Gb and 60 to 80 million reads (Liferechnologies, 2012, Merriman et al., 2012, ThermoFisher, 2014), which is a great advancement in high-throughput and results in significant time and cost reduction.

With regard to calculating the scalability that affects the number of samples and the choice of the chip used, a calculation by Ion community was suggested to guide the determination of the number of samples that can be used for the chip. It involves the length of the target to be sequenced, read length and coverage depth. For example, if the sample target size is 20 kb and the coverage needed is 20x, (20 x 20) = 400 kb total reads would be obtained from each sample. Therefore, for this example, Ion 314^{TM} chip, which has the lowest capacity (20 Mb) for a 200-base read, can sequence 50 samples simultaneously (which resulted from the calculation 20Mb/400kb=50) (http://ioncommunity.lifetechnologies.com/message/15695#15695).

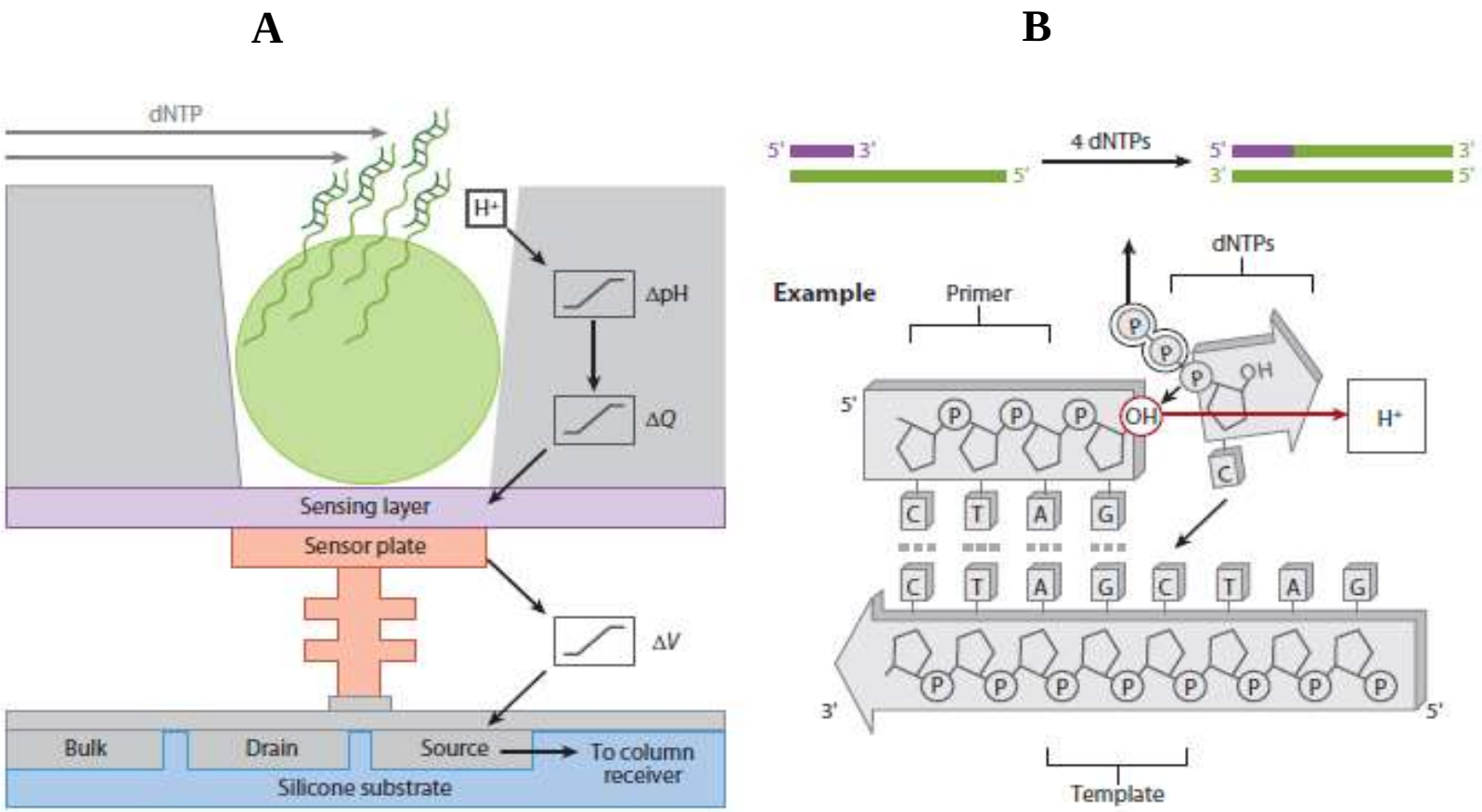


Figure 1.6 Ion PGM™ sequencing technology and chemistry.

A) The structure of the well in the chip is shown with the templated bead. A flow of nucleotides allows incorporation to correspondent sequence that leads to a hydrogen ion H^+ per nucleotide. This changes the pH in the solution that is detected by the sensor underneath the well, converting the information into voltage that becomes digitalised to be readable base calling. (B) the sequencing by syntheses is illustrated, in which nucleotide addition leads to H^+ release. (Adapted from (Mardis, 2013)

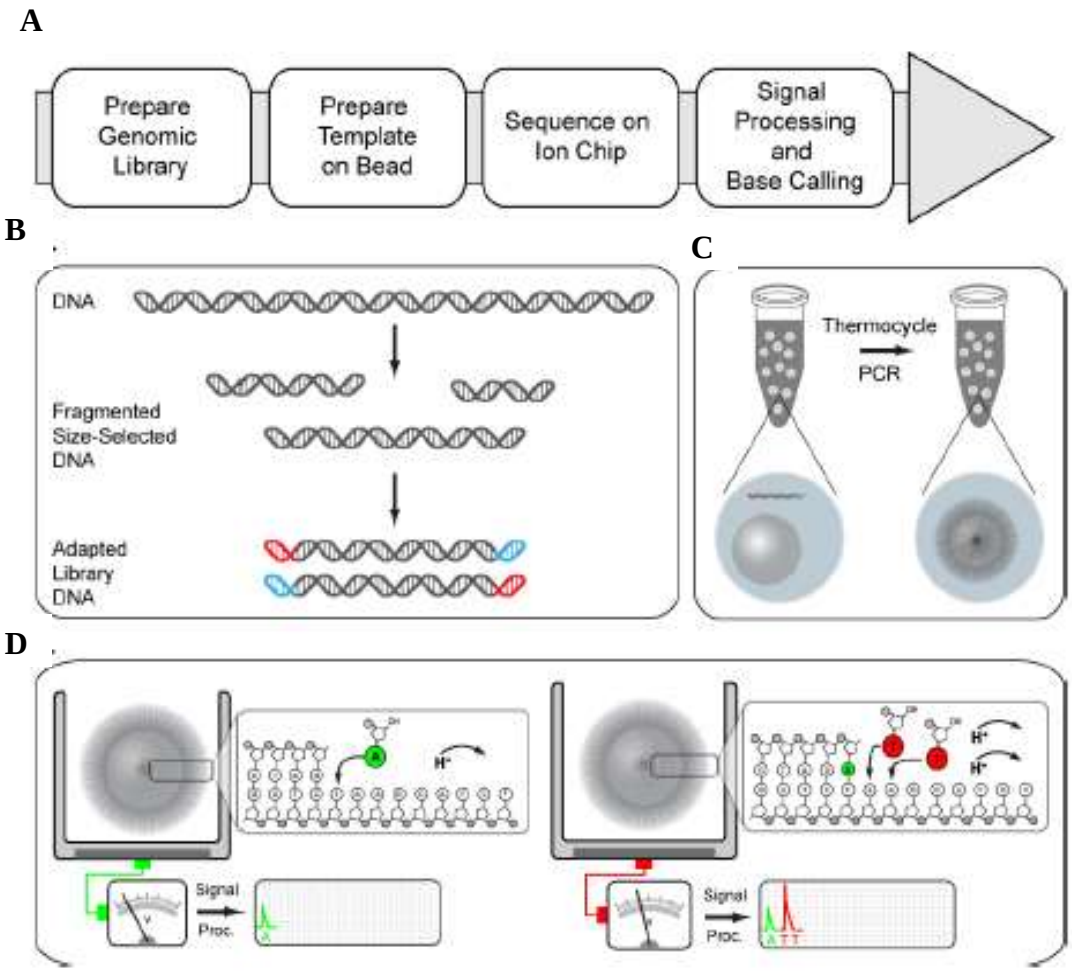


Figure 1.7 Summary of the Ion PGM™ workflow.

A) An overview for the steps of Ion PGM™ workflow from library preparation to sequencing results. (B) DNA library fragmentation and ligation before, (C) clonal amplification of the single fragment attached to a bead, by emulsion PCR. (D) Beads are then loaded into wells in the chip, in which the sequencing process and the base calling commence. The incorporation of one nucleotide gives one signal, while incorporation of 2 nucleotides gives a doubled (higher) signal and so on. (Adapted from (Rothberg et al., 2011).

Table 1.4 Updated chips v2 of Ion PGM™ in terms of capacity and run time according to read length

(ThermoFisher, 2015).

	Read length	Ion 314™ Chip v2	Ion 316™ Chip v2	Ion 318™ Chip v2
Output	200 base	30–50 Mb	300–600 Mb	600 Mb–1 Gb
	400 base	60–100 Mb	600 Mb–1 Gb	1.2–2 Gb
Run time	200 base	2.3 hr	3.0 hr	4.4 hr
	400 base	3.7 hr	4.9 hr	7.3 hr

1.8.7 NGS and BGG

Several groups have found the benefits and feasibility of utilising NGS platforms in the genotyping of blood groups. Stabentheiner and colleagues (2011) used GS FLX (NGS platform) to sequence samples with serologically weak Rh phenotypes and found that the majority of polymorphisms for *RHD* were comparable (95%) to results of Sanger sequencing (Stabentheiner et al., 2011). Rieneck at al. (2013) used Illumina GAIIx to successfully predict the phenotype of foetal Kell blood group status from cell-free foetal DNA in maternal plasma (Rieneck et al., 2013). Another group illustrated that NGS is suitable for genotyping blood group genes; 18 genes involved in 15 blood group systems were genotyped for the clinically significant antigens by using (PGM™) with AmpliSeq™ technology (Fichou et al., 2014). The latter involves a custom designed panel, in which multiplex primers are provided, resulting in amplicons that target areas

of interest that are mainly the coding and untranslated regions (Merriman et al., 2012). Prediction of the blood group phenotype from the database containing the data produced by the NGS platforms was described in 2015 (Giollo et al., 2015). Subsequently, another group illustrated the prediction of the phenotype for RBCs antigens and platelets from whole genome sequencing completed by Illumina HiSeq NGS platform (Lane et al., 2016). Furthermore, another group show the feasibility of high throughput genotyping of ABO blood group by NGS platform (Illumina) that enabled the studies of the allele frequency and discovery of novel alleles (Lang et al., 2016). Additionally, NGS coupled with LR-PCR approach has been used recently to genotype blood group genes *RHD*, *RHCE* and *KEL* (Halawani, 2015).

1.9 Thesis aims

BGG provides more comprehensive blood group typing than the conventional serology. BGG with the high-throughput capability allows large-scale genotyping for extended blood groups in individuals (donors and recipients). This would have a significant impact in reducing the risk of alloimmunisation, especially in those in need of multiple transfusions, such as those with sickle cell disease (SCD). NGS allows a sequencing-based genotyping with high-throughput that provides large scale genotyping in a discovery mode, allowing the definition of all polymorphisms including novel and rare ones.

In this project, the powerful NGS-based genotyping platform was used to provide comprehensive genotyping of the blood group genes *FY*, *JK* and *ABO*. The Ion PGM™ NGS platform with Long range PCR (LR-PCR) was used to sequence the entire gene plus flanking regions in each case. Accordingly, all polymorphic mechanisms can be explored. In general, the feasibility of using NGS for genotyping these genes will be determined. The sequence-based mode of the NGS would allow the discovery of all polymorphisms, including rare and novel ones, which would provide comprehensive analysis of the molecular basis of the blood group systems alleles, thus more accurate reflection and prediction of the phenotype from alleles.

The objectives for this project were:

FY blood group system

- To develop and optimise a reliable NGS protocol for extensive *FY* gene genotyping.
- To explore all the polymorphisms in the coding regions (exons).
- Also, to explore the polymorphisms in other regions, such as introns, splice sites and regulatory regions. This would provide more informative reflection on the

phenotype (predicted phenotype).

- To analyse the association of intronic polymorphisms with *FY* alleles.
- To use FY BGG by NGS as a test platform for BGG of other blood group genes using NGS

JK blood group system

- To develop and optimise a reliable NGS protocol for high resolution *JK* gene genotyping.
- To utilise the discovery mode of the NGS to detect all existing polymorphisms in the exons.
- To assess the polymorphisms in introns, splice sites and flanking regions.
- To analyse the correlation of detected polymorphisms with the key *JK* SNPs to evaluate their allele specificity, allowing discovery of allele specific sequences.

ABO blood group system

- To develop and optimise a reliable NGS protocol for comprehensive *ABO* gene genotyping.
- To analyse all polymorphisms in all 7 exons, in addition to those in introns, splice sites and outer regions.
- To study the correlation of intronic polymorphisms with the *ABO* alleles.

CHAPTER 2

Materials and Methods

2.1 Genomic DNA (gDNA) and RNA from whole blood samples

RNA and genomic DNA (gDNA) were extracted and purified from whole blood samples collected (with consent and full ethical approval) from randomly selected anonymous donors obtained from National Health Service Blood and Transplant (NHSBT; Filton, Bristol UK)). Ethylenediamine tetraacetic acid (EDTA) tubes containing whole blood samples were centrifuged at 3500xg for 5 minutes in order to separate the whole blood component and obtain the buffy coats (containing white blood cells), which were then processed to extract DNA. RNA was extracted from the first layer of red blood cells (RBCs) of the supernatant, in which most of the reticulocytes (nucleated red cells) are believed to be found (refer to Chapter 4 of *JK*).

2.2 Sequencing of blood group genes by next-generation sequencing

The current study uses targeted region sequencing using the Ion PGM^{TM} sequencer (Thermo Fischer Scientific Inc., Leicestershire, UK), by which the entire gene, including up- and downstream flanking regions (in order to discover all polymorphisms within and outside of the gene exon-intron, such as those in promoter regions) are sequenced. The first step of the process is preparation of a genomic library (containing gDNA); this involves targeting the region by long-range polymerase chain reaction (PCR), in which long PCR overlapping products (about 4-14kb amplicons) are generated from gDNA, utilising specific primers that cover the entire target gene plus flanking regions, without missing parts of the gene sequence. Following this, the long amplicons are chemically fragmented by enzymes (or physically by sonication), resulting in fragments of approximately 200bp in size. These fragments will be then ligated with barcoded adapters that are used to distinguish between samples. The next

step involves template preparation, in which the ligated fragments (including individual fragments) are immobilised on beads via complementary base pairing with the adapter sequence, and undergo clonal amplification by use of emulsion PCR (emPCR). Each step in the described procedure is followed by purification and quality control (QC) steps using magnetic bead purification and Bioanalyzer®, respectively. Next, the enriched beads are loaded onto an ion chip, within which the sequencing commences when inserted into the Ion PGM™ (Thermo Fischer Scientific Inc., Leicestershire, UK). Finally, the data will be analysed and aligned with the sequence of a reference gene of interest, using various bioinformatics software packages which interpret polymorphisms by referring to specific database. See (Figure 2.1.) for a summary of the workflow.

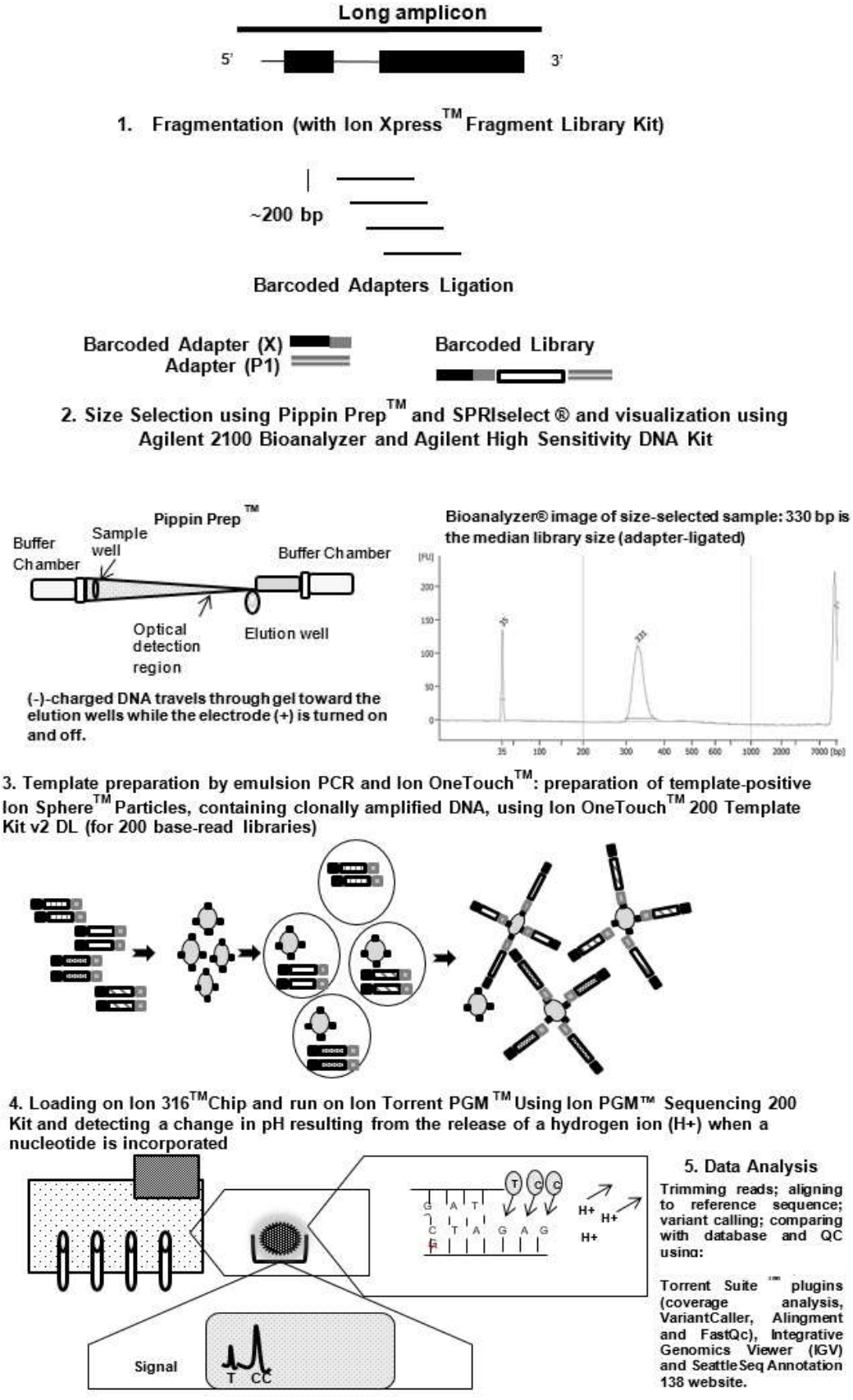


Figure 2.1. Overview of the process for sequencing blood group genes by Next-Generation Sequencing. The main steps involve gDNA library construction, template preparation, sequencing and data analysis.

2.2.1 gDNA extraction and purification

Sample gDNA was extracted using a QIAamp DNA Blood Mini Kit (QIAGEN, Hilden, Germany), which is based on a spin column procedure extracting and purifying DNA from the buffy coat. The kit contains the following: buffers (lysis buffer AL and washing buffers AW1 and AW2 with 96-100% ethanol added) and lyophilised QIAGEN Protease, dissolved in 1.2ml protease solvent (preservative).

Whole blood samples were transferred into 15 ml Falcon tubes and centrifuged at 2500xg for 10 minutes at room temperature (20^{o}C) to obtain three distinctive layers: plasma; buffy coat (containing concentrated nucleated leukocytes) and the RBC-concentrated layer. The buffy coat was collected in 1.5ml tubes and spun at 10000 rpm to obtain a distinctly separated buffy coat from the extra serum and red cells, following which 200µl of the buffy coat was transferred into a new 1.5ml tube and placed on ice to be used in the procedure. The DNA extraction procedure followed: 20µl QIAGEN Protease was thoroughly mixed with 200µl buffy coat and an equal volume (200µl) of AL buffer by pulse-vortexing for 15 seconds, until a homogenous solution was observed. Subsequently, tubes were briefly centrifuged and then incubated at 56^{o}C for 10 minutes (suggested to give the best DNA yield), followed by a pulse spin to collect residual sample form the inside of the tube lids. 200 µl ethanol (96-100%) was added to the samples and mixed by pulse-vortex for 15 seconds, then spun briefly. This mixture was then carefully applied to the QIAamp mini spin column in the provided 2ml collection tube and centrifuged at 16300xg (to avoid clogging of buffy coat) for 1 minute, before transferring the column to new 2ml tubes and discarding the collection tube that contains the filtrate.

Subsequently, 500µl Buffer AW1 was added to the mixture and spun for 1 minute at 16300xg and the above column centrifugation process was repeated. Following this,

500µl Buffer AW2 was added to the QIAamp mini spin column and centrifuged at 20,000xg for 3 minutes. Afterwards, the collection tubes were replaced with new 2ml Eppendorf ® tubes to perform further centrifugation at 20,000xg for 1 minute, as recommended by the company, in order to remove any residual Buffer AW2. The QIAamp mini spin column was then placed into new sterile 1.5ml Eppendorf ® tubes and the filtrate discarded. Following this, 200µl Buffer AE were added (for the purpose of DNA elution) to the column and incubated for 5 minutes (to increase DNA yield) at room temperature. Samples were then centrifuged at 16300xg for 1 minute. DNA was divided in two separate sterile aliquots and stored at - 20^{o}C. DNA was quantified using several approaches and platforms, including NanoDrop^{TM} (Thermo Fischer Scientific Inc., Leicestershire, UK), NanoVue^{TM} Plus (GE Healthcare Life Sciences, Buckinghamshire, UK) and Qubit® 2.0 Fluorometer with Qubit® dsDNA Broad range (BR) assay Kit (Invitrogen^{TM}, Paisley, UK).

2.2.2 RNA extraction and purification

Samples designated for RNA extraction were processed immediately by centrifugation at 2500xg for 10 minutes at room temperature in order to separate the whole blood components, plasma, buffy coat, and red cells (Figure 2.2). Subsequently, the buffy coat was carefully transferred into sterile 1.5 ml Eppendorf ® tubes and placed on ice for gDNA extraction; however, RNA extraction was conducted first due to its higher vulnerability.

RNA extraction and purification were accomplished using QIAamp RNA Blood Mini Kit (QIAGEN, Hilden, Germany), with addition of RNase-Free DNase Set (QIAGEN, Hilden, Germany). The top 1 ml of RBC layer (thought to be the location of the reticulocytes) was transferred into sterile 1.5 ml tubes and mixed with 5x the sample volume (5 ml) EL buffer, incubated for 15 minutes on ice and briefly vortexed twice to obtain a transparent solution (indicating lysis of erythrocytes). Next, samples were spun

at 400xg for 10 minutes at 4^{o}C before removing the supernatant carefully, leaving a trace of the supernatant (100µl) - advised for RNA extraction from reticulocytes (instead of leucocytes, for which the kit was intended by the manufacturers). 2 ml Buffer EL was added to the samples, and cells were re-suspended by vortexing briefly, followed by centrifugation at 400xg for 10 minutes at 4^{o}C. Afterwards, the supernatant was discarded carefully, leaving a small trace (50µl) which was mixed with 600µl RLT buffer containing 143 mM β-mercaptoethanol (β-ME) (Sigma-Aldrich Company Ltd, UK), then vortexed and transferred directly onto a QIAshredder spin column, placed in 2 ml collection tubes. Samples were then spun for 2 minutes at maximum speed to obtain homogenised lysates, while columns were discarded. 600 µl 70% ethanol was added to the homogenised lysates and mixed by pipetting, then carefully transferred into new QIAamp spin column tubes. The columns were then centrifuged at ≥8000xg for 15 seconds, following which the flow-through and collection tubes were discarded and replaced with new collection tubes. Subsequently, an on-column DNase digestion was conducted by washing the spin column membrane, via the addition of 350µl of Buffer RW1 to the column, and spun at ≥8000xg for 15 seconds. Again, the flow-through and collection tubes were discarded and replaced with new collection tubes. Next, 80µl incubation mix was prepared by adding 10µl DNase I stock solution to 70µl Buffer RDD and mixed gently (due to the sensitivity of DNase to physical denaturation) by inverting the tube. This was followed by brief spinning of the solution and its direct addition to the QIAamp spin column membrane (to ensure a full DNase digestion) for incubation at room temperature for 15 minutes. Next, 350µl Buffer RW1 was added to the column; the samples were centrifuged for 15 seconds at ≥8000xg and the flow-through discarded. The columns were then transferred into provided 2ml collection tubes, followed by the addition of 500µl working solution of Buffer RPE mixed with 4 volumes of 96-100%) ethanol, and then spun again for 15 seconds at ≥8000xg

(discarding the collection tubes containing the flow-through. This step was repeated; however, samples were then centrifuged at 20,000xg for 3 minutes. In order to prevent Buffer RPE carryover, columns were then transferred into new 2 ml Eppendorf ® tubes and re-spun at 20,000xg for 1 minute. For RNA elution, the QIAamp spin columns were transferred into provided 1.5 ml micro centrifuge tubes and 40µl RNase-free water was directly added into the column membrane, followed by centrifugation at ≥8000xg for 2 minutes. 2µl RNA was used for quantification by NanoDrop 2000™ (Thermo Scientific, USA).

2.2.2.1 RNA quantification

Quantification by Nano Drop 2000™ (Thermo Scientific, USA) involved blanking the spectrophotometer with 1µl nuclease free water and loading 1µl RNA, resulting in a concentration of ng/ µl and the purity of the RNA from contaminants (given by the absorbance ratio of A_{260}/A_{280}) at the recommended range of 1.9-2.

In turn, gDNA was extracted from samples as explained in Section 2.2.1 and quantified by Qubit® 2.0 Fluorometer with Qubit® dsDNA Broad range (BR) assay Kit (Invitrogen™, Paisley, UK).

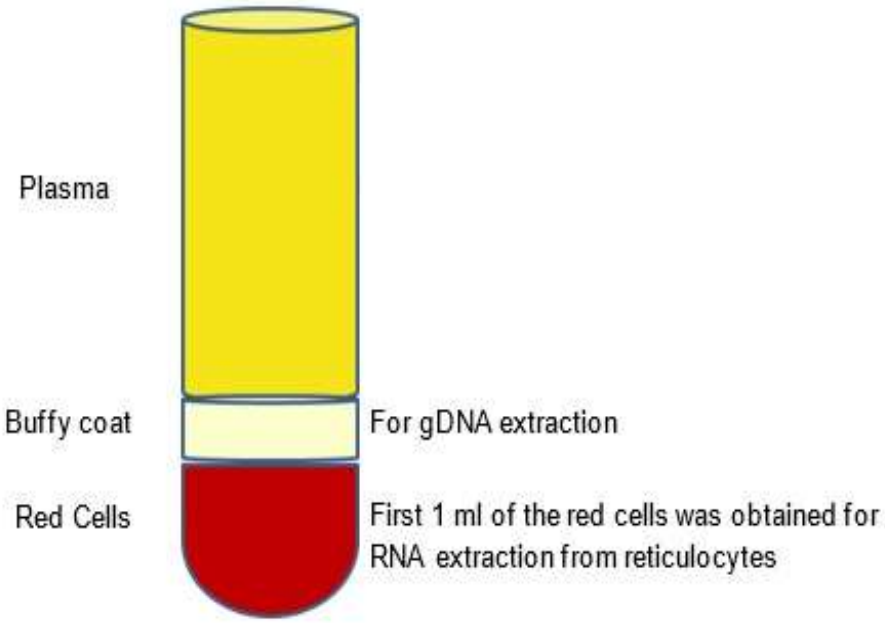


Figure 2.2. Separation of blood components in a tube.

2.2.3 DNA quantitation

The PCR amplicon quantitation is very critical for the downstream processing; therefore, it is crucial to quantify the gDNA and PCR amplicons as accurately as possible. In the very first gDNA quantification attempt of this project, NanoDrop™ (Thermo Scientific, USA), NanoVue™ (GE Healthcare Life Sciences, Buckinghamshire, UK) and NanoDrop 2000™ (Thermo Scientific, USA) were used due to availability. However, comparison of data with that obtained using a Qubit® 2.0 Fluorometer and Qubit® dsDNA Broad range (BR) assay Kit (Invitrogen™, Paisley, UK) showed that the former approach (apart from NanoDrop 2000™) resulted in clear overestimation of the concentration of extracted gDNA. Data obtained from NanoDrop 2000™ were comparable, although slightly elevated, to that of Qubit® 2.0 Fluorometer (data not shown). Therefore, the Qubit® 2.0 Fluorometer, which is more accurate due to its high selectivity for double-stranded DNA over RNA, was subsequently used. There are two kits for DNA quantification: Qubit® dsDNA Broad range (BR) and High sensitivity

(HS) assay Kit, both of which are used with the Qubit® 2.0 Fluorometer to measure DNA concentration. The former is capable of measuring concentrations within the range of 100pg/µl to 1000ng/µl, whereas the latter measures within a range of 10 pg/µl to 100 ng/µl. Therefore, gDNA quantification was conducted using the Broad range (BR) assay Kit, which consists of Qubit® dsDNA Broad Range Reagent, buffer, standard 1 and standard 2. First, all reagents were brought up to room temperature (22-28 ^{o}C) for 30 minutes (or until complete thawing of Qubit® (BR) reagent), according to manufacturer instructions. Then, Qubit® working solution was prepared as follows: a 1:200 dilution of Qubit® reagent in Qubit® BR buffer was achieved by adding 1 x n µl reagent to 199 x n µl buffer (where n = the number of samples plus the 2 standards), in a sterile 50 ml tube although a volume for 5 extra samples were added to avoid pipetting errors. Subsequently, 2µl from of extracted gDNA and 10µl of each standard were added to 0.5ml Qubit® assay tubes (Invitrogen^{TM}, Paisley, UK), and a final volume of 200µl per tube was achieved by adding 198µl of working solution to the samples tubes and 190 µl to the standard tubes. Next, the tubes were vortexed for 3 seconds, followed by incubation at room temperature for 2 minutes in order to reach the optimal fluorescence of the Qubit® assay. Before measuring the sample tubes, the Qubit® 2.0 Fluorometer was calibrated by measuring the standards, following which data was obtained as a Qubit® 2.0 Fluorometer (QF) value. Concentrations of samples (gDNA) were calculated as follows: {Sample concentration = QF value x (200/2)} in ng/ µl.

2.2.4 Amplicon library preparation for next-generation sequencing

2.2.4.1 Primers

Several software packages were utilised to design the oligonucleotide primers for *FY*, *JK* and *ABO*. Primer 3 software (http://frodo.wi.mit.edu/primer3/) was initially used to design the primers; the NCBI BLAST database (http://blast.ncbi.nlm.nih.gov/) was used to confirm the specificity of primers; the UCSC Genome Browser database

(http://genome.ucsc.edu) was also used to analyse the specificity of the primers; in addition, visualisation of the location of the amplicons within the genome was done using the in-silico PCR (virtual PCR) service in UCSC.

Primers were designed to target the entire *FY*, *JK* and *ABO* genes, including the flanking (upstream and downstream) regions, by generating long amplicons. For the *FY* gene, the entire gene and flanking regions were targeted by single Long-range PCR amplicon, due to the small size of the *FY* gene (3194bp according to the Genecards database (www.genecard.org; Table 2.1). For *JK* and *ABO* genes, several amplicons had to be designed to cover the larger *JK* (28765bp) and *ABO* (20039bp) genes (3 and 4 amplicons for *JK* and *ABO*, respectively). The amplicons on the latter two genes (*JK* and *ABO*) overlapped and were of various sizes (Tables 2.2 and 2.3). With regard to amplifying *ABO*, two amplicons (named 1A, adapted from (Huh et al., 2011), and 1B), covering the upstream region, exon 1 and part of intron 1, were used since not all of *ABO* alleles were amplified with the same primer pair. Nevertheless, the other 3 amplicons were utilised for all samples (Table 2.3). The primers were received lyophilised from Eurofins Genomics (Ebersberg, Germany) in HPLC (High Performance Liquid Chromatography) purity, and were re-suspended by adding nuclease-free water (Ambion®, Applied Biosystems, Thermo Fisher Scientific, USA**)** as indicated by the manufacturer.

Table 2.1. LR-PCR forward (F) and reverse (R) primers for *FY* amplification. Each primer is 23 bp in size.

	Primers (5'- 3')	Amplicon size	Position within the human genome 19
F	GTGTGAGTGAGTGAGAGGCAGAG	4784bp	chr1:159171901+159176684
R	GCCAGAGAGGAGACAGAAGACAG		

Table 2.2. LR-PCR forward (F) and reverse (R) primers for *JK* amplifications. Each primer is 25 bp in size (the combined size of amplicons is 36730bp).

Amplicon		Primers (5'-3')	Amplicon size	Position within the human genome
1	F	GAAGCCCACTGCGAAATCCAAATAG	11012bp	chr18:43301432+43312443 covering upstream, exons 1-5
	R	TGAGGGCAAATGGGAGGTGATACAA		
2	F	GCTTTACCTCATCCCTTCCAGACAA	11053bp	chr18:43310959+43322011 cover exons:5-9
	R	GCTTCTGCCCTCTATTGTAACACTC		
3	F	GCTTTGGGTCTCTGGCTTTAGTGTA	14665bp	chr18:43318623+43333287 cover exons 8-11
	R	TTCCGTGCTAATCCTGTATCATGGG		

Table 2.3. LR-PCR of forward (F) and reverse (R) primers for *ABO* amplifications (the combined size of amplicons, with 1A or 1B, is 35430bp or 36187bp, respectively). The length of the primers is given in brackets after the primer sequence.

Amplicon		Primers (5'-3')	Amplicon size	Position within the human genome
1A	F	CTTACCAAAGGAGTCACACCCTCAAA (26)	6260bp	chr9:136149487 -136155746 upstream, exon1 and part of intron 1
	R	GAACTCAGCGATATTGAACACAGTGC (26)		
1B	F	CTCAAACGATCACATTCAATGCTTGC (26)	7017bp	chr9:136148978 -136155994 upstream, exon1 and part of intron 1
	R	CACTCATTTGCCAGGTTTCTCAAAGA (26)		
2	F	GGCCCTCAGGTGATATAGGAGTTAAG (26)	9188bp	chr9:136140844 -136150031 intron 1 (connecting amplicon 2 with 3)
	R	AATAGAAGGATGGTTCGCAGAGACTT (26)		
3	F	AAAACTGCAATATAAACCAGGCACCA (26)	10354bp	chr9:136132466 -136142819 part of intron1 to part of intron 7
	R	CAGGAAACATCTGGAGCCTTGTATTG (26)		

4	F	GCTGAGCTAACCTTGGGAGACATTT (25)	9628bp	chr9:136124605 +136134232 part of intron 4, exon 7 and downstream
	R	CATCCAGGTTGTACCAAGTGTCAGA (25)		

2.2.4.2 LR-PCR amplicon amplification

In a UV-sterilised hood, 50µl (PCR) master mix for LR-PCR reactions was prepared in 0.2 ml sterilised nuclease free tube strips (BIOplastics, Landgraaf, the Netherlands). The master mix contained: 1x LongAmp® Hot Start Taq Master Mix (New England BioLabs Inc., Herts, UK); 500nM forward and reverse primers; and gDNA template (100ng for *FY*, as shorter than the other genes, 200ng for *JK* amplicons 1 and 2, 400ng for amplicon 3 and 200ng for *ABO*). Sterile filter tips (Fisher Scientific, Loughborough, UK) and sterile forceps were used to handle tubes and minimise contamination. First, gDNA was added first to the 0.2 ml tubes; PCR mix, prepared separately in 1.5 ml Eppendorf ® tubes, was then added and samples were mixed gently by pipetting up and down, followed by brief spinning prior to thermocycling. Thermocycling was carried out on a Veriti Thermal Cycler (Life Technologies, Paisley, UK) and included the following optimisation attempts (data not shown): temperature gradients for annealing and the other stages; and time alteration. The optimised thermocycling conditions for *FY*, *JK* and *ABO* (amplicons 2, 3 and 4) are described in Tables 2.4, 2.5 and 2.6. Following gDNA amplification, amplicons were visualised by agarose gel electrophoresis.

The *ABO* amplicons 1A and 1B master mix contained: 200ng gDNA; 1x Phusion Flash High-Fidelity PCR Master Mix (Thermo Scientific, Leicestershire, UK); and 1000nM forward and reverse primers. The master mix for *ABO* amplicons 2, 3 and 4 contained: 200ng gDNA, 1x LongAmp® Hot Start Taq Master Mix (New England BioLabs Inc,

Herts, UK); and 400nM forward and reverse primers. The thermocycling conditions can be seen in Table 2.7.

Table 2.4. Optimised thermocycling conditions for *FY*

Step	Temperature	Time	Cycles
Initial denaturation	95^{o}C	30 seconds	1
Denaturation	95^{o}C	30 seconds	30
Annealing	62^{o}C	30 seconds	
Extension	72^{o}C	3 minutes	
Final extension	72^{o}C	5 minutes	1
Holding	4^{o}C	∞	-

Table 2.5. Optimised thermocycling conditions for *JK*

Step	Temperature	Time	Cycles
Initial denaturation	94^{o}C	5 minutes	1
Denaturation	94^{o}C	30 seconds	30
Annealing	60^{o}C	30 seconds	
Extension	65^{o}C	10 minutes	
Final extension	65^{o}C	10minutes	1
Holding	4^{o}C	∞	-

Table 2.6. Optimised thermocycling conditions for *ABO* amplicons 2, 3 and 4.

Step	Temperature	Time	Cycles
Initial denaturation	94°C	5 minutes	1
Denaturation	94°C	30 seconds	30
Annealing	62°C	30 seconds	
Extension	65°C	10 minutes	
Final extension	65°C	10minutes	1
Holding	4°C	∞	-

Table 2.7. Optimised thermocycling conditions for *ABO* amplicons 1A and 1B. *Two-step PCR (annealing and extension combined) was used as the Tm value was approximately 72°C when calculated by Thermo Scientific's Tm calculator.

Step	Temperature	Time	Cycles
Initial denaturation	98°C	3 minutes	1
Denaturation	98°C	10 seconds	35
Extension*	72°C	2 minutes	35
Final extension	72°C	3 minutes	1
Holding	4°C	∞	-

2.2.4.3 Agarose gel electrophoresis

Agarose gel electrophoresis was used to separate gDNA fragments by size and visualise the presence of the amplicon amplified by the PCR reaction. For *FY* and *ABO* samples, 1% Hi-Res Standard Agarose (Geneflow Limited, Staffordshire, UK) was prepared by melting 1.4g of the agarose powder in 140ml of 1xTAE (40mM Tris-acetate-1mM EDTA) and staining with 14µl GelRed (10,000X, BT41003, Cambridge Bioscience). 3µl Generuler 1kb Plus DNA ladder (Thermo Scientific, Leicestershire, UK) was used

to assess the average size of the amplicons (Figure 2.3). For each sample, 8µl PCR reaction mix was mixed with 2µl DNA gel loading buffer (containing 100mM Tris–HCl, pH 8.3, 10% (*v/v*) glycerol and 0.05% (*w/v*) Orange G dye), 10µl of which was then loaded into individual wells. DNA was electrophoretically separated at 85 V for 1 hour for *FY* and at 80 Volts for 1 hour and 40 minutes for *ABO*. The same process was carried out for *JK* analysis; however, the first two amplicons were separated by 1% agarose gel, whereas the third amplicon was separated by a 0.8% (w/v) agarose gel, according to its size (approximately 15kb), and run for 1 hour and 20 minutes at 70 V. Amplicons were then prepared for purification.

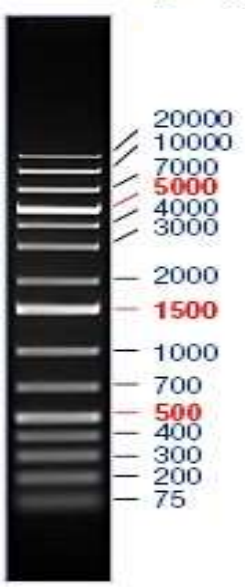


Figure 2.3. GeneRuler^{TM} 1Kb Plus DNA ladder (Thermo Fisher Scientific) used as a marker of DNA size

2.2.4.4 Purification

After checking for the presence of the amplicons, the remaining PCR mix (40 µl) for each sample was purified using magnetic beads purification, which ensures removal of primer dimers and free nucleotides that may interfere with downstream processing. At the start of the project, particularly for the first two sequencing experiments of *FY* and *JK*, Agencourt® AMPure® XP beads Reagent (Beckman Coulter, High Wycombe, UK) was used for the purification of the PCR product, as well as for all purification steps in the gDNA library preparation process (Figure 2.4). However, SPRIselect® reagent kit (Beckman Coulter, High Wycombe, UK), which utilises the same principle, was used for subsequent experiments. SPRIselect® reagent kit (Beckman Coulter, UK) has an advantage over the AMPure® XP beads Reagent, in that the former can be used for a size selection approach (Section 2.2.4.10), in addition to as a means of purification for the library. Other advantages of using SPRIselect® reagent kit for purification and size selection include reduced time, errors, costs and increased throughput of libraries, when compared to using Pippin Prep™ instrument (Sage Science, Inc., Beverly, USA) for size selection (Section 2.2.4.10). According to manufacturer's instructions, beads were re-suspended at room temperature for 30 minutes before commencing the purification (this step is only required for AMPure® XP beads in contrast to those of SPRIselect® reagent kit). 70% molecular grade-ethanol (Fisher Scientific, UK) diluted in nuclease-free water (Ambion®, Applied Biosystems, Thermo Fisher Scientific, USA**)** was used for washing steps.

In a nuclease-free 1.2ml well plate (Thermo Scientific Abgene, USA), 40µl PCR product was mixed with 72µl (1.8x the sample volume) bead solution by pipetting and incubated at room temperature for 7 minutes. This particular bead-sample ratio is suggested to attract a PCR product of >100bp and eliminate contaminants, such as free nucleotides and primer dimer which are of smaller sizes (Figure 2.5). The plates

containing PCR product were then placed on a magnetic plate (Agencourt® SPRIPlate 96R - Ring Super Magnet Plate, Beckman Coulter, UK) for 3 minutes, which promoted attachment of amplicon-carrying beads to the side of the wells, leaving a clear supernatant on top, which was then carefully discarded. Following this, without removing the plate from the magnet, two washing steps were carried out: 30µl freshly prepared 70% ethanol was added to wells and incubated for 30 seconds at room temperature, before the supernatant was carefully discarded without disturbing the pellet. Afterwards, without removing the plate from the magnet, residual ethanol was removed by pipetting with smaller size pipette tips (20µl; to avoid disturbing the pellet) and the beads air dried for 3 minutes at room temperature.

Next, purified amplicons were eluted by removing the plate from the magnet plate and re-suspending the pellet by adding 15µl of nuclease-free water (Ambion®, Applied Biosystems, Thermo Fisher Scientific, USA**).** This was followed by a thorough trituration of the solution (10x; to release the DNA into the water), after which plates were placed back onto magnetic plates until the solution became clear. The solution, containing purified DNA (amplicons), was then transferred to new 0.2 ml sterilised nuclease free tube strips (BIOplastics, Landgraaf, the Netherlands), with 2µl used for amplicon quantification. The PCR amplicons were first quantified using the Qubit® 2.0 Fluorometer with High sensitivity (HS) assay kit; however, as the DNA concentration of some samples was higher than 100 ng/µl, the Qubit® dsDNA Broad range (BR) assay Kit (Invitrogen™, Paisley, UK) was subsequently used, as it measures higher concentration samples more accurately. In addition, it is more cost-effective since it is recommended that extracted gDNA to be quantified by a BR kit. Amplicon quantification is critical for downstream stages of the library construction process, such as amplicon pooling, fragmentation and ligation, in order to result in full coverage of the target.

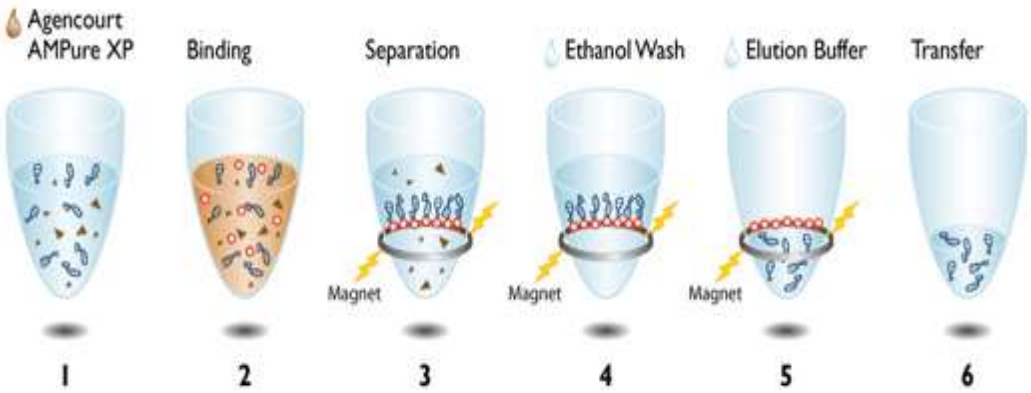


Figure 2.4. The principle of the purification mechanism using magnetic beads (applies to both AMPure® XP and SPRIselect® beads reagent kits). 1) Beads are added to the PCR reaction, 2) magnetic beads bind PCR amplicons (the desired size range determines the bead-amplicon ratio), 3) separation of PCR amplicons from contaminants using magnetic beads, 4) ethanol wash, 5) elution of PCR amplicons from the magnetic particles and 6) transfer of amplicons from beads into new tube strips (Beckman Coulter Inc., 2015).

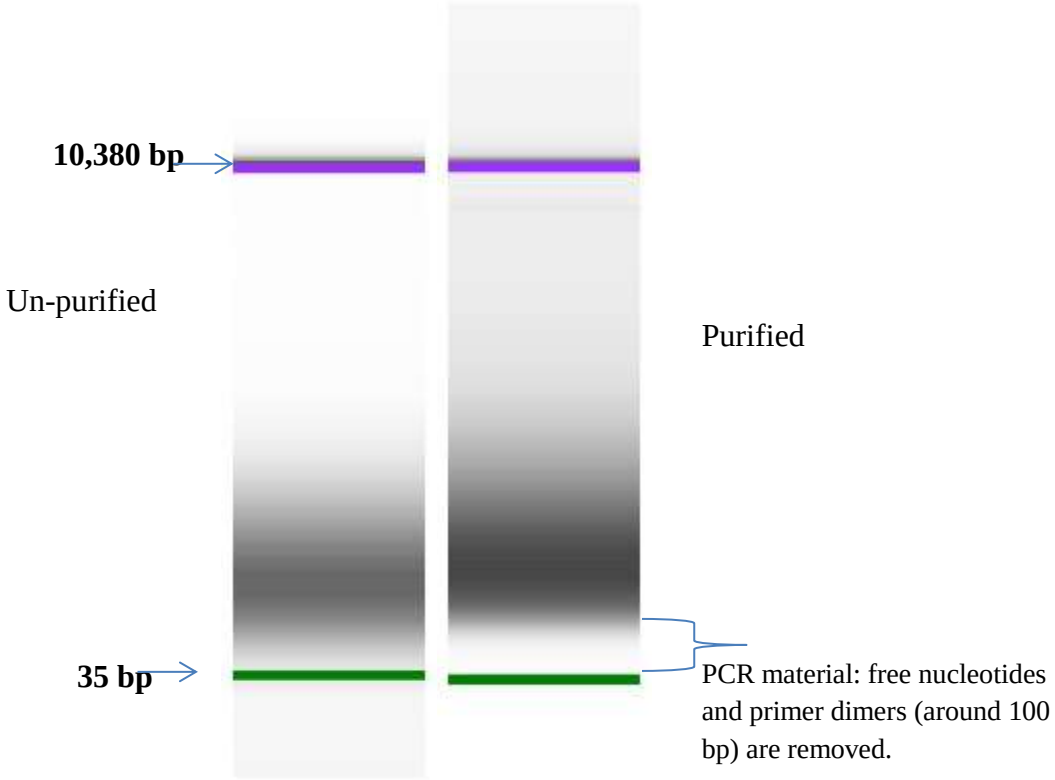


Figure 2.5. Example Bioanalyzer® readout of purified and un-purified samples. Upper and lower DNA markers are shown.

2.2.4.5 Sample aliquots

In order to prepare for the next step of the process (amplicon library fragmentation), 100ng DNA was required, according to manufacturer's instructions. Accordingly, based on the concentrations obtained from the Qubit® 2.0 Fluorometer, equal concentrations from each amplicon (which ensures equal representation of each amplicon) were calculated to provide 100ng, when pooled. First, aliquots of equal concentrations were prepared to simplify the pooling process, by adding equal volumes from each sample of amplicon and avoiding pipetting errors. With regard to *FY* samples, aliquots containing 20 ng/µl DNA in 30µl nuclease-free water (Ambion®, Applied Biosystems, Thermo Fisher Scientific, USA**)** were prepared from each sample for the purpose of DNA fragmentation. A few samples containing low DNA concentrations were diluted in 15µl

nuclease-free water. This was done using the following equation: C1 (from Qubit®) x V1 (volume taken from the stock) = C2 (20ng/µl) x V2 (30µl or 15 µl). As a result, 5 µl (= 100ng) from each aliquot was added to the fragmentation mix.

Unlike *FY*, which has only one amplicon, the complete *JK* gene target is represented by 3 amplicons, plus flanking regions, which therefore needed to be pooled in order to be fragmented together. As 100ng was required from 3 combined amplicons, 33.3ng (100/3 = 33.3ng) of each amplicon was added to the fragmentation reaction. First, aliquots were prepared to achieve a concentration of 16.6ng/µl (33.3/2) in a total volume of 10µl (using a similar equation to that used for *FY*); therefore, 2µl of each amplicon aliquot provided, in total, the 100ng needed for the fragmentation reaction. Similarly, 25ng (100ng/4) of each of the 4 amplicons of *ABO* required pooling to provide 100ng for the fragmentation reaction mix. Aliquots of DNA at a concentration of 12.5ng/µl were prepared in a total volume of 10µl, applying a similar equation to that used for *FY* and *JK*; therefore, 2µl of each amplicon was added to the fragmentation reaction to provide 100ng. Regarding aliquot preparation, a number of samples with low DNA concentrations, that did not require dilution with nuclease-free water (Ambion®, Applied Biosystems, Thermo Fisher Scientific, USA**)**, were instead directly added to the fragmentation reaction mix (volumes chosen resulted by dividing 33.3 by the *JK* amplicon concentration and 25 by the *ABO* amplicon concentration.

2.2.4.6 Amplicon library fragmentation (purified fragmented library)

For this step, Ion Xpress™ Plus Fragment Library Kit was used, which includes two kits: Ion Shear™ Plus Fragment library Kit (for fragmentation) and Ion Plus Fragment Library Kit (Life Technologies, Paisley, UK). The samples were enzymatically fragmented (sheared) using Ion Shear™ enzyme and Ion Shear™ Plus Fragmentation Kit, which contains: Ion Shear™ Plus 10x reaction buffer, Ion Shear™ Plus Enzyme Mix II and Shear™ Plus stop buffer. Before preparing the fragmentation reaction mix,

Ion Shear™ Plus 10x reaction buffer and Ion Shear™ Plus Enzyme Mix II were thoroughly mixed by vortexing and brief centrifugation, followed by placing on ice (according to manufacturer's instructions). In sterile 1.5ml Eppendorf LoBind® tubes, the following contents were added in the order recommended by the manufacturer: 100ng pooled amplicons (for *FY* amplicons, 5µl were added and 2 µl of *JK* and *ABO* amplicons were added, with the exception of low concentration samples); 5µl Ion Shear™ Plus 10x reaction buffer (pre-mixed prior to addition) plus nuclease-free water (Ambion®, Applied Biosystems, Thermo Fisher Scientific, USA**)** to make up the total reaction volume to 40µl. The fragmentation mix was then briefly vortexed (for 5 seconds) and pulse-spun. Subsequently, 10 µl Ion Shear™ Plus Enzyme Mix II (pre-mixed prior to addition) was added to the mix to reach a total volume of 50µl, which was then rapidly triturated (10x), avoiding bubbles. The mixing step, using a 40µl pipette tip, ensures homogenous mixing of enzyme with amplicons and buffer. Afterwards, the tubes were incubated at 37^{o}C for the manufacturer-recommended amount of time that yields a median fragment size of 200-300bp (acquiring a target read length of 200 bases, which is suitable for downstream preparation steps, such as ligation and template preparation). After several optimisation attempts (see below), the incubation length for the reaction time was set for 5 minutes for *FY* samples and 8 minutes for *JK* and *ABO (*the variation in time is due to the difference in PCR product sizes). Next, 5µl Shear™ Plus stop buffer was immediately added and mixed vigorously by vortexing for 10 seconds, following which the tubes were placed on ice.

The reaction tubes were purified by Agencourt® AMPure® XP beads Reagent in the first two experiments with *FY* and *JK* but in subsequent experiments were purified using the SPRIselect® reagent kit (Beckman Coulter, UK). For the purification, the 50µl reaction mix was transferred to nuclease-free 1.2 ml well plates (Thermo Scientific, USA), then thoroughly mixed with 90µl (1.8 x sample volume) magnetic bead solution

by trituration (8-10x) to homogenise the solution, followed by incubation at room temperature for 7 minutes. Subsequently, the steps of the bead purification were the same as explained in Section 2.2.4.4, with the exception of the volume of 70% ethanol used for washes, which here was 500 µl (2 washes). Following air drying of beads, the plate was removed from the magnetic plate (Agencourt® SPRIPlate 96R - Ring Super Magnet Plate, Beckman Coulter, UK) and the fragmented library was eluted in 25µl Low (10mM Tris pH 8.0, 0.1mM EDTA) (TE), which is provided in the Ion Plus Fragment Kit (Life Technologies, Paisley, UK), by the same method of elution explained in Section 2.2.4.4 To check for shearing integrity, 2µl of purified supernatant (containing purified sample) was tested by Agilent® 2100 Bioanalyzer® instrument and Agilent High Sensitivity DNA Kit (Agilent Technologies Ltd, UK) (Section 2.2.4.7), while the rest of purified fragmented library (PF) was transferred to new 0.2 ml sterilised nuclease free tube strips (BIOplastics, Landgraaf, the Netherlands) in preparation for the next step (adapter ligation, Section 2.2.4.9).

2.2.4.7 The Agilent® 2100 Bioanalyzer

The Bioanalyzer™ (Agilent Technologies, California, USA) is an on-chip electrophoresis system that provides digital data and information regarding the size distribution, quantitation and quality control of DNA, RNA, proteins and cells, which are displayed as gel-like images (bands) and electropherograms (peaks; Figure 2.6). For this project, Agilent High Sensitivity DNA Kit (Agilent Technologies UK Limited) was used for the Agilent® 2100 Bioanalyzer® instrument, which is capable of accurately analysing a size range of 50-7000bp. The Agilent High Sensitivity DNA Kit contains: 10 high-sensitivity DNA chips (11 samples per chip can be analysed), a syringe kit, electrode cleaner and spin filters. The reagent box contains: high-sensitivity DNA ladder, high-sensitivity DNA marker (lower 35bp and higher 10380bp marker); high-sensitivity DNA dye concentrate; and gel matrix. Prior to chip preparation and

experimentation, a few adjustments were made to the chip priming station, which is where the chip is placed while the samples are loaded. With each new kit, the syringe was replaced in priming solution to prevent an air leak and maintain optimum pressure. Then, the base plate of the priming station was adjusted to the *C* marker, while the syringe clip was adjusted to the lowest position. Subsequently, Gel-Dye mix was prepared by, first, allowing the DNA dye concentrate and the gel matrix to equilibrate to room temperature for 30 minutes in the dark, to completely thaw the DMSO and protect the gel-Dye from light. Then, the high sensitivity DNA dye was vortexed for 10 seconds and briefly centrifuged, following which 15 µl was added to the high-sensitivity DNA gel matrix. The gel-dye mix was then vortexed for 10 seconds and carefully transferred to the spin filter (provided within a collection tube), followed by a spin for 10 minutes at room temperature at a speed of 2240xg ± 20%, (equal to ~2680xg). The filter was discarded and the mix was kept for future use, protected from light, for 6 weeks as per manufacturer's instructions. Next, the gel-dye mix (equilibrated to room temperature) was loaded onto the high-sensitivity DNA chip and placed on the priming station. 9 µl gel-dye mix was dispensed carefully to the bottom of the well (indicated in Figure 2.7), taking special care to avoid pipetting error. The syringe plunger was then positioned to 1ml before carefully closing the priming station, the station properly closed (as indicated by a click). Next, the plunger was pressed down for 60 seconds and released by the clip release mechanism (to reach the 0.3 ml), following which the plunger was briefly left for 5 seconds and then slowly pulled back to the 1ml mark. Subsequently, the chip priming station was opened and 9µl of the gel-dye mix was added to the remaining 3 wells (Figure 2.7), followed by the addition of 5µl of high-sensitivity DNA marker to all 11 sample wells plus the ladder well. Afterwards, 1µl DNA ladder was loaded into the indicated well, and 1µl of each of the 11 samples was added to the relevant sample wells. 1µl DNA marker was also added to any unused sample wells

(Figure 2.7). Then, the chip was placed horizontally onto the adapter of the IKA vortex mixer (Model MS3, supplied together with the Bioanalyzer® instrument) and vortexed at 2400 rpm for 60 seconds. Within the next 5 minutes, the chip was placed on the 2100 Bioanalyzer® instrument and the readout was carried out, which took 45 minutes. During this, the electrodes were cleaned with the electrode cleaner chip. The name of the software used was the Agillent 2100 Expert software.

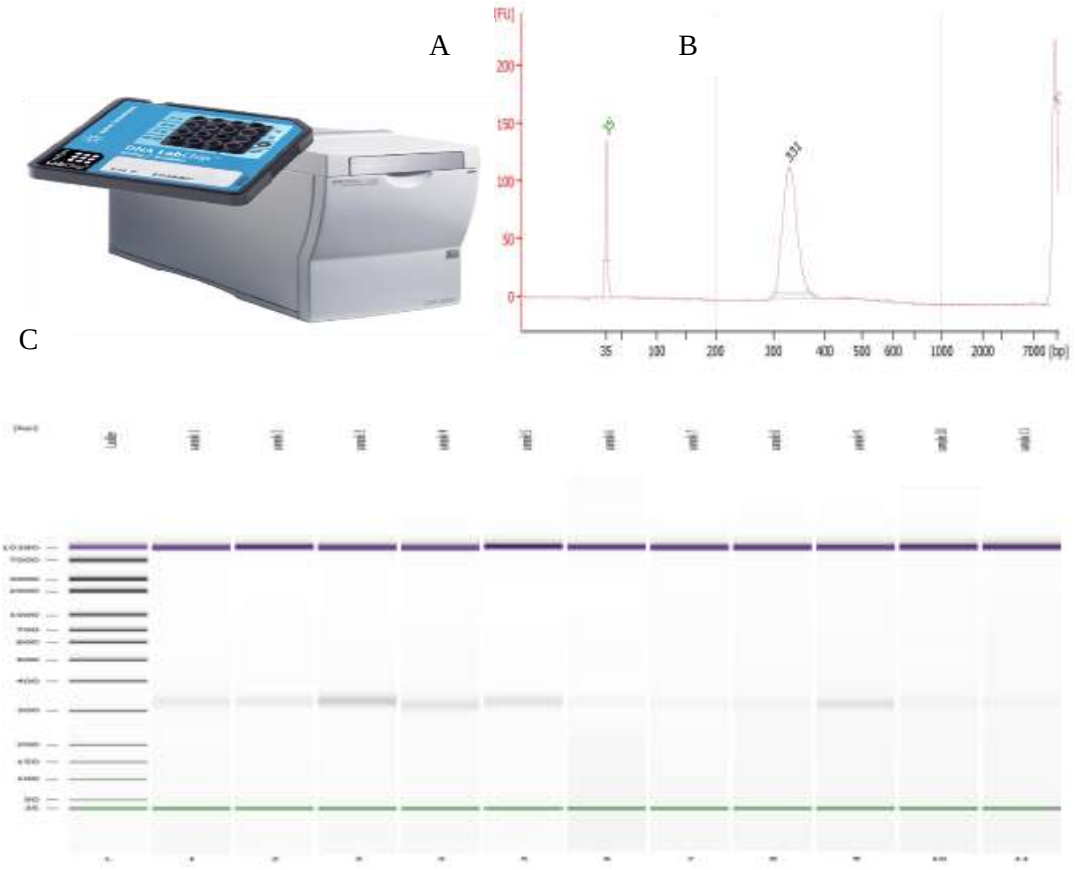


Figure 2.6. The Bioanalyzer® instrument and the chip, into which samples are loaded, B) an example of data showing peaks and sizes of samples (X axis shows size and signal intensity is in Y) and C) the gel-like bands (Agilent Technologies, California, USA).

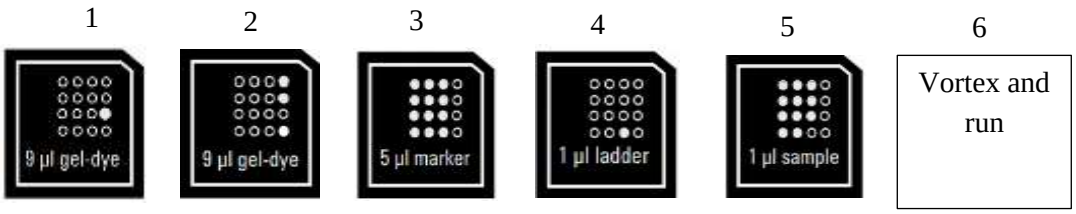


Figure 2.7. The wells of the High sensitivity DNA chip, into which samples are loaded before running on the Bioanalyzer® instrument.

The indicated wells (white) of the high-sensitivity DNA chip. 1) the well that is to be pressurized. 2) 9 gel-dye mix was added to the indicated wells. 3) 5 µl DNA marker was added to all 12 wells, including the one containing the DNA ladder. 4) 1 µl DNA ladder was carefully added. 5) Samples were added to the indicated wells, while 1 µl DNA marker was added where sample was not needed. 6) The chip was then vortexed for 60 seconds at 2400 rpm, prior to its placement into the Bioanalyzer® instrument for the readout (adapted from Agilent Technologies, California, USA).

2.2.4.8 Optimisations of the fragmentation reaction time (to confirm the desired fragment size range)

Although it is recommended in the protocol that a 15 minute incubation may be required for the reaction to yield a 200-300bp median fragment, with a fragment size range of 100-700 bp, the optimum reaction (fragmentation) was obtained following several optimisations that included testing of different time periods. In addition, fragmentation size distribution is not only affected by the incubation time, but also is very sensitive to practical handling, such as temperature variation and vortexing time. The Agilent® 2100 Bioanalyzer® instrument and Agilent High Sensitivity DNA Kit (Agilent Technologies, California, USA) were chosen to test the various incubation times as well as the shearing integrity. The examination of several incubation times for *FY* were: 20, 15, 10 and 5 minutes, the last of which was chosen for better size

distribution (Figure 2.8), as there is a difference in size distribution and median fragment size between samples processed for 20 minutes and 5 minutes.

For *JK*, incubation on the heating block was tested at 10, 15 and 20 minutes. As 10 and 15 minutes were acceptable, since the size of pooled amplicons (~36kb) that was larger than *FY*, when checked by Bioanalyzer ®instrument, an incubation time of 12 minutes was chosen for the first *JK* sequencing experiment. Then, samples from the fragmented library were purified, as explained in Section 2.2.4.4 and then checked by Agilent® 2100 Bioanalyzer® instrument and Agilent High Sensitivity DNA Kit (Agilent Technologies, California, USA). Although samples fragmented for 12 minutes were eventually sequenced successfully, they did not achieve the recommended parameter for optimum output; therefore, in subsequent experiments for *JK*, samples were fragmented for 8 minutes, which gave a better size range (Figure 2.9). As the total size for the pooled amplicons of *ABO* were similar to that of *JK* (~36kb), fewer optimisations were carried out and included test incubation periods of 6 and 7 minutes (Figure 2.10). Following these tests, 8 minutes was eventually used for *ABO* sample fragmentation.

FY fragmentation time

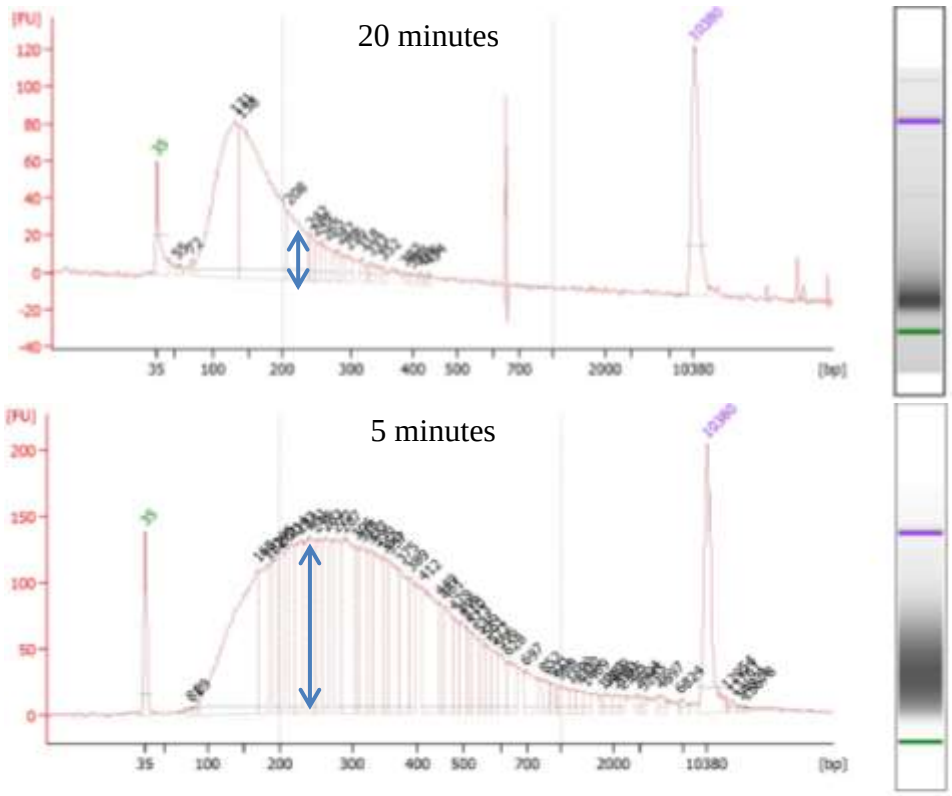


Figure 2.8. Variations in the peak of 200-300bp between *FY* samples fragmented for 20 or 5 minutes. Samples processed for 20 minute show that the majority of fragments are approximate 150 bp in size, which is lower than the required median concentration of the library. Samples fragmented for 5 minutes yielded better concentrations of the library, at around 200-300 bp, which is the required range. Results shown were obtained using the Bioanalyzer® instrument.

JK fragmentation time

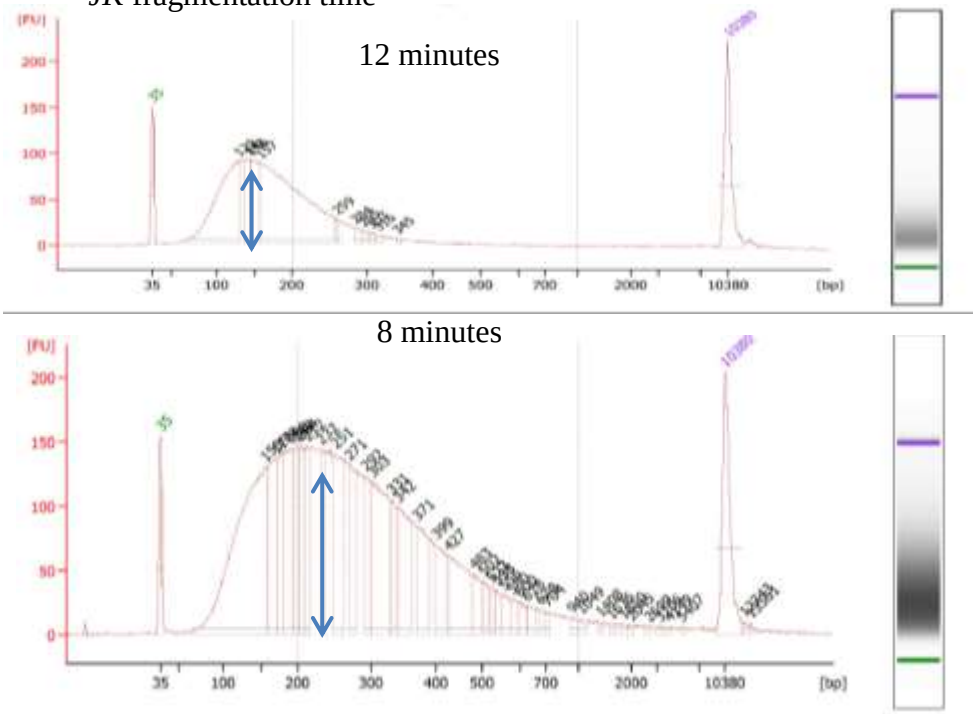


Figure 2.9. Variations in the peak of 200-300bp between *JK* samples fragmented for 12 or 8 minutes. Samples processed for 12 minutes showed peaks approximating 150 bp in size, which is lower than the required median concentration of the library. Samples fragmented for 8 minutes yielded better concentrations of the library, at around 200-300 bp, which is the required the range. Results shown were obtained using the Bioanalyzer® instrument.

ABO fragmentation time

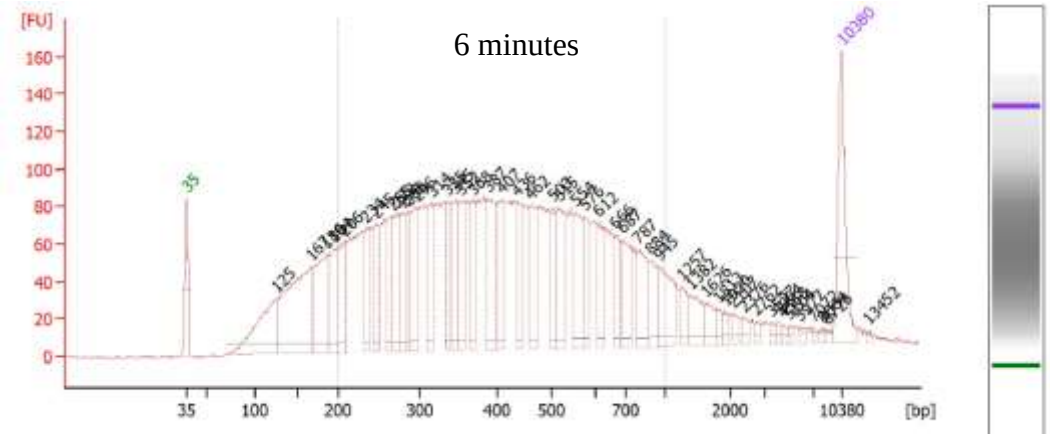


Figure 2.10. The size distribution of an *ABO* sample fragmented at 6 minutes. There is a good distribution of 100-700 bp and a good peak at 200-300bp. However, 8 minutes was the incubation period used for subsequent *ABO* samples in the experiments due to the size similarity with *JK*, which would yield more concentrated fragments at around 200-300bp. Results shown were obtained using the Bioanalyzer® instrument.

2.2.4.9 Ligation of barcoded adapters (purified-ligated library)

Following fragmentation, and ensuring a size range of 100-700bp with optimum target median fragment size of around 200-300bp, the samples were ligated to barcoded adapters (P1 and Ion Xpress™ Barcode X adapter) from the Ion Xpress™ Barcode adapters Kit (Life Technologies, Paisley, UK). P1 is the universal adapter that is compatible and recognised by the Ion PGM™, and the X adapter (X referring to the desired barcode number) contains a distinct barcode (about 13bp) to distinguish the samples when pooled, prior to sequencing. The combined size of the adapters is ~80bp.

For general ligation, the kit used was the Ion Plus Fragment Library Kit (Life Technologies, Paisley, UK) that contains: 5x End repair buffer, End repair enzyme (which is only used if the samples are fragmented physically), 10x Ligase buffer, DNA Ligase, Nick repair polymerase, dNTP mix, adapters (which are used if non-barcoded adapters are required), Platinum® PCR SuperMix High Fidelity, Library Amplification Primer Mix (which is used with Platinum® PCR SuperMix High Fidelity, in case the sample library needs amplification afterwards), and Low TE. The ligation of barcoded adapters involved transferring the fragmented purified library (PF; 20µl) to 0.2ml sterilised nuclease-free tube strips (BIOplastics, Landgraaf, the Netherlands) from the previous step of fragmentation. Then, each of the following reagents were triturated before their addition to the 0.2ml tube in the indicated order: 10 µl 10x Ligase buffer; 2µl universal adapter (P1); 2µl distinctive Ion Xpress™ Barcode X adapters (taking care to avoid cross-contamination or mixing adapters); 2 µl dNTP mix; 54µl nuclease-free water (Ambion®, Applied Biosystems, Thermo Fisher Scientific, USA); 2µl DNA ligase; and 8µl Nick repair polymerase, which led to a total reaction mix volume of 100µl. The reaction mix was then triturated using 80µl pipette tips (taking care to avoid bubbles) and briefly spun. Afterwards, the reaction was placed on a Veriti Thermal Cycler (PCR thermocycler; Life Technologies, Paisley, UK) to start the nick repair reaction, by

which the linkage between the adapters and the DNA fragments is completed. The thermocycling conditions were: 25^{o}C for 15 minutes; 72^{o}C for 5 minutes; and 4^{o}C for up to 1 hour, as this was not stopping point. Next, within 1 hour, the reaction mix (100µl) was transferred to fresh nuclease-free 1.2ml well plates (Thermo Scientific Abgene, USA) to prepare them for the next step of bead purification. 120µl (1.2x sample volume) of SPRIselect® reagent was mixed thoroughly by trituration until a homogenous solution was visible, followed by incubation at room temperature for 7 minutes. The plate was then placed on a magnetic plate (Agencourt® SPRIPlate 96R - Ring Super Magnet Plate, Beckman Coulter, UK) for 3 minutes, or until the solution became clear, following which the supernatant was carefully removed without disturbing the pellet. The purification process then continued, as in Section 2.2.4.4, including two wash with 500µl 70% ethanol (freshly prepared) up to the elution step where 20µl low TE was added, after which the plate was removed from the magnet. Next, the samples with low TE were vigorously mixed by trituration (at least 10 times), before being placed back on the magnet and incubated for 1 minute or until a clear solution was observed as the magnet beads migrated to the bottom of the plate. Subsequently, 2µl of reaction mix was used to check the size difference, which was expected to increase by about 80bp (Figure 2.11A and B), using the Agilent® 2100 Bioanalyzer® instrument and Agilent High Sensitivity DNA Kit (Agilent Technologies UK Limited), while the rest of the PL samples (18µl) were transferred to new 0.2ml sterilised nuclease-free tube strips (BIOplastics, Landgraaf, the Netherlands) and stored at - 20 ^{o}C for subsequent use. A few sequencing batch samples proceeded directly to the size selection step without freezing.

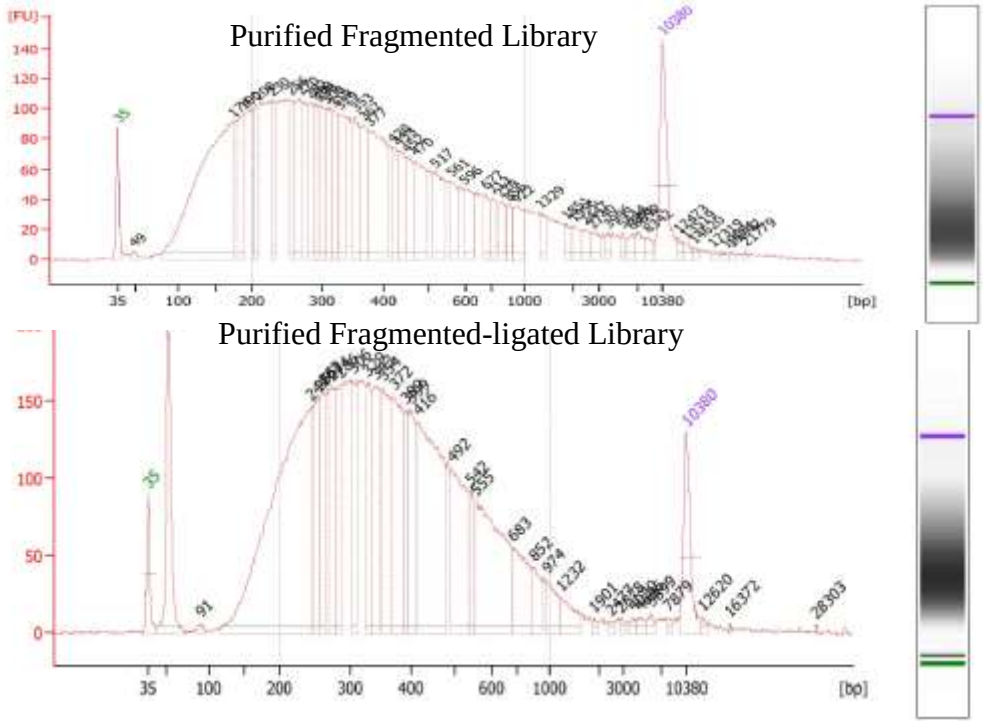


Figure 2.11.A A sample of the *JK* fragmented library (top graph) and the same library with adapter ligated (bottom graph) resulting in shifting of the size (increasing) due to the addition of adapters. Images shown were obtained from the Bioanalyzer® instrument.

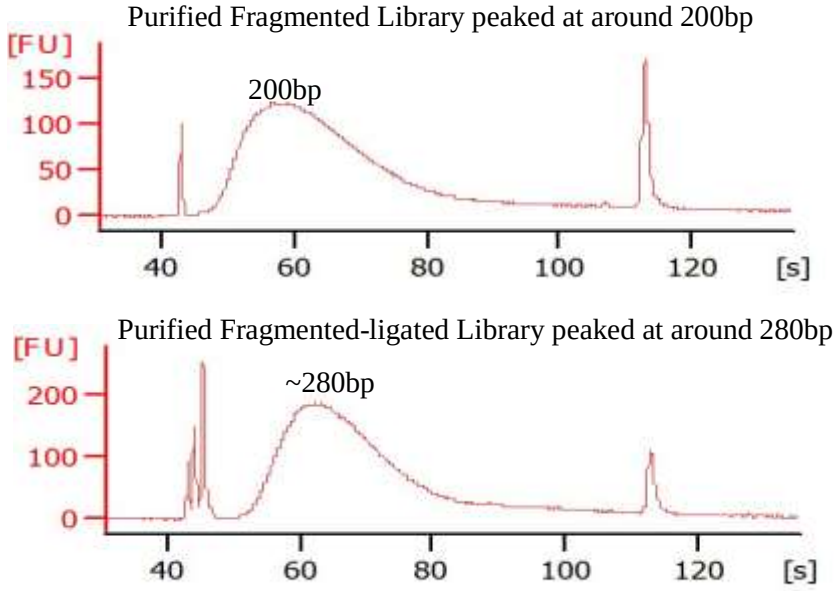


Figure 2.11.B A sample of *FY* fragmented library (top graph) and the same library with adapter ligated (bottom graph), resulting in a ~80bp increase in size. The X axis is seconds, as the ladder required manual alteration to give bp reading. Images shown are obtained from the Bioanalyzer® instrument.

2.2.4.10 Size selection of the library

In order to achieve the optimum DNA target read length, the adapter-ligated library was size selected for optimum length, which was conducted for the first two experiments of *FY* and *JK*, using the Pippin Prep™ instrument (Sage Science, Inc.**,** Beverly, USA) and Pippin Prep™ Kit 2010 with Ethidium Bromide cassettes (Figure 2.12). Although the Pippin Prep™ instrument was used for size selection of the first two experiments of *FY* and *JK*, SPRIselect® reagent kit (Beckman Coulter, UK) was subsequently used for several reasons. First, SPRIselect® reagent kit (Beckman Coulter, UK) has an advantage in terms of multi-functionality and cost effectiveness, as in the library construction protocol it can be used for both purification and size selection (negating the need for the former after size selection), in contrast to the Pippin Prep™, which needs a purification step using AMPure® XP beads Reagent. Second, although the Pippin Prep™ provides a narrow and more consistent size range, which can be altered as required (tight or broad; Figure 2.13), it is time consuming, expensive and labour-intensive. This is because each of the Pippin Prep™ Kit 2010 with Ethidium Bromide cassettes (10 cassettes per kit) is only capable of processing 4 samples (Figure 2.12) per run, which takes about 1 hour and 10 minutes (for a peak at ~330bp). In contrast, significantly more samples can be processed by a 60ml SPRIselect® reagent kit (Beckman Coulter, UK), with less time required to accomplish the size selection procedure. Since the Pippin Prep™ was only used for the first experiments, the protocol is listed here briefly. Pippin prep is automated electrophoresis with fluorescence-based DNA detection that can be programmed to separate the DNA by size.

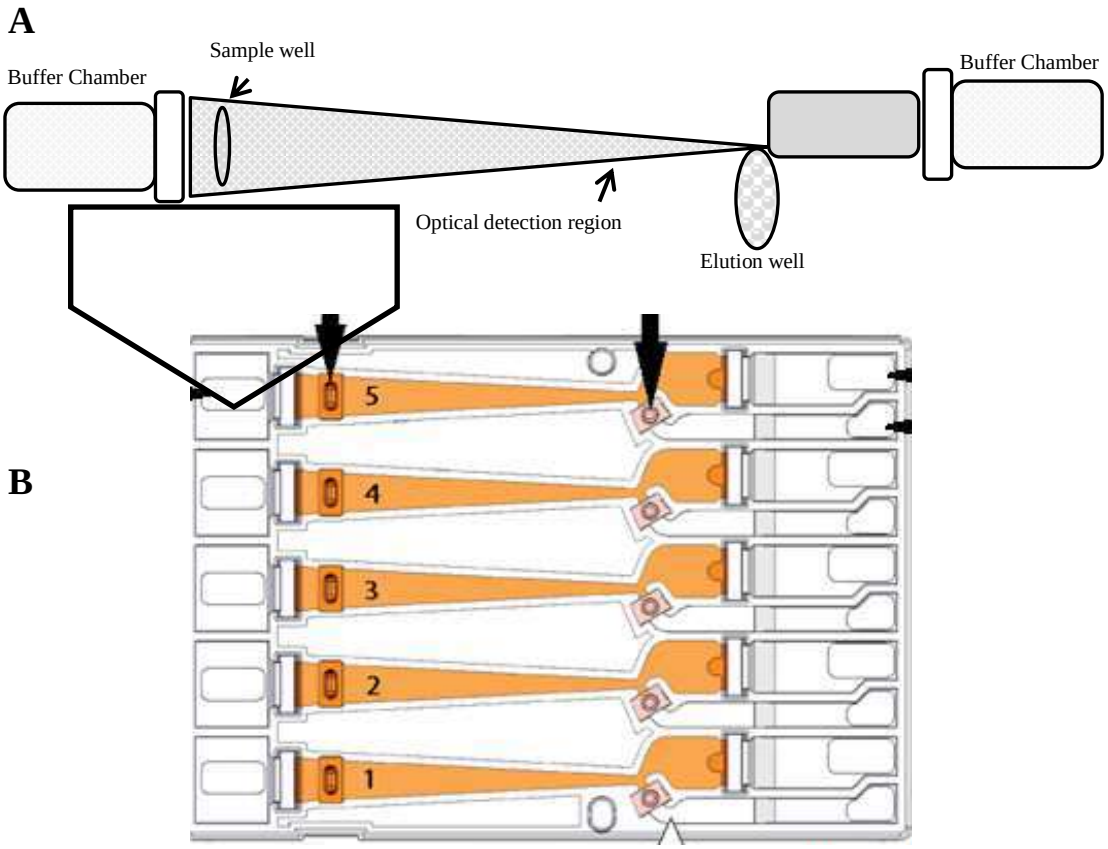


Figure 2.12. The Pippin Prep™ Kit 2010 with Ethidium Bromide cassette. A) a zoom in one of the 4 samples lanes in the gel-filled cassette, through which the DNA travels toward the elution wells at the positive electrode. B) The Pippin Prep™ cassette which has only 4 lanes for samples. NOTE: 4 lanes are used for samples while the fifth is for the marker.

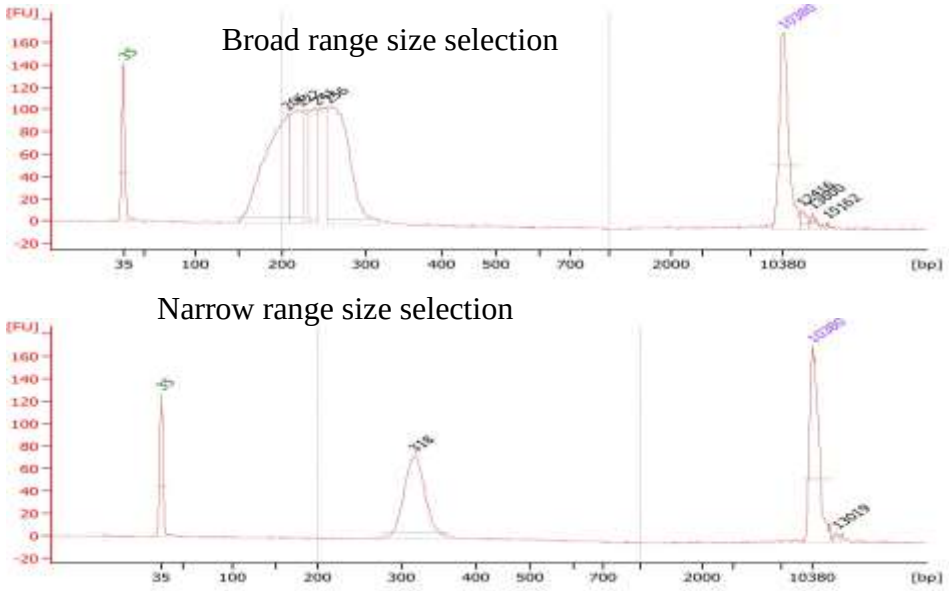


Figure 2.13. Two *JK* samples from the first experiment. (Top) experiment programmed to be size selected for broad range (peak around 200-250bp), which elevates the yield of the size selected DNA according to the protocol. (Bottom) the sample was size selected for a narrow range (peak at 318bp). All samples were processed by the Pippin Prep^{TM} instrument.

In terms of using the Pippin prep™ instrument, first the machine optics were calibrated using the provided plate, as instructed, by entering 0.8 in the LED target setting I ph, mA. Then, the cassette to be used was inspected visually for any leakage and bubbles behind the elution wells, which can be removed by tapping and removing the adhesive strips to place in the optical nest. Next, all the original buffer from the elution wells was replaced by fresh 40µl electrophoresis buffer, after which the wells were sealed using adhesive strips. The level of buffer in the sample wells was filled to the top before running the continuity test, which measures the current in each cassette channel (the separation-elution). For a successful run, this test requires a PASS result before running the samples. Before loading the samples, the PL samples were prepared for the Pippin prep™; to achieve this, the volume was made up to 30µl by adding 12µl low TE and then mixing with 10µl of provided loading solution, pre-equilibrated to room temperature. Then, the mix was vortexed and spun briefly, before commencing the loading procedure. This entailed removing 40µl of buffer from the first sample well (which was assigned as the marker well) and immediately replacing it with 40µl of provided marker B. This loading process was repeated for the remaining 4 samples. The loading was carried out carefully to avoid damaging the gel by piercing the bottom or the sides of the wells. Although the recommended protocol for library preparation involved a certain programme (BP target setting at 315bp for library size of 200bp), the size selection range programme selected was mainly broad to give a higher yield (Tables 2.8.A 2.8.B). This programme was optimised and selected according to size distribution of the fragmented-purified samples. After the run stopped (90 minutes), samples (~40-60 µl; some made up to 60µl by adding nuclease-free water (Ambion®, Applied Biosystems, Thermo Fisher Scientific, USA) to unify the beads ratio) which were in Tris-TAPS buffer (50 mM Tris, 30 mM TAPS, 0.1 mM EDTA) were collected in 1.5 ml Eppendorf ® tubes, followed by purification (Section 2.2.4.4) with 1.8x the

volume beads (Agencourt® AMPure® XP), washes with 70% ethanol, and elution with 25 µl low TE. Subsequently, the samples were checked by the 2100 Bioanalyzer® instrument for sample distribution and concentration. However, the first 12 size-selected samples for *FY* showed very low distribution (around 200-300bp), which accounted for the low concentrations of the size selected library. Therefore, an extra step of amplification of the library was carried out to increase the size selected library and the accuracy of the 2100 Bioanalyzer® instrument, as well as to determine the dilution factor for the purpose of template preparation.

For the amplification, 20µl of adapter-ligated and size selected library was mixed with 100µl Platinum® PCR SuperMix High Fidelity and Library Amplification Primer Mix (provided in the Ion Plus Fragment Library Kit), then run on the thermal cycler with the following PCR programme: initial denaturation at 95^{o}C for 5 minutes; then (for 5 cycles instead of 8 in order to avoid PCR errors) denaturation at 95 ^{o}C for 15 seconds; annealing at 58 ^{o}C for 15 seconds; extension at 70 ^{o}C for 1 minute; and hold at 4^{o}C. Then, the amplified library was purified by the bead-sample ratio of 1.5X the volume of the sample and eluted in 20µl Low TE, before being stored at -20^{o}C ready to be checked by the 2100 Bioanalyzer® instrument.

Table 2.8.A. The size selection range programme for *FY* samples. The column titles are exactly as shown on the instrument screen. BP target is the expected peak size; BP start is where the fragment with this size is collected; BP ends is where the collection is stopped. *It is considered that the broader the range the higher amount of DNA collected, while smaller amounts of DNA are collected in tight ranges. The inconsistent ranges were selected during optimisations, where it was noticed that those ranges were slightly comparable. Each cassette can take up to 4 samples and a control. Run time was ~1-2 hours.

Sample	BP Target	BP start	BP ends	Description of range*
1b	322	290	355	Broad
2	315	290	340	Tight
3	322	290	355	Broad
4	320	290	350	Broad
5	318	290	350	Broad

Table 2.8.B. The size selection range programme for *JK* samples. The column titles are exactly as displayed on the instrument screen. BP target is the expected peak size; BP start is where the fragment with this size is collected; and BP ends is where the sample collection stops. *****It is suggested that the broader the range the higher amount of DNA collected, while in tight ranges smaller amounts of DNA are collected. The inconsistent ranges resulted during optimisations, where it was noticed that those ranges were slightly comparable. Each cassette can take up to 4 samples and a control. Run time was ~1-2 hours. Samples 1, 2, 4 and 5 had a very broad range to result in higher yield, as those samples where fragmented for longer time.

Sample	BP Target	BP start	BP ends	Description of range*
1	240	170	310	Broad
2	240	170	310	Broad
3	320	290	350	Broad
4	240	170	310	Broad
5	235	160	310	Broad

Size selection with the SPRIselect® reagent kit (Beckman Coulter, UK) was started by transferring 18µl PL samples into nuclease-free 1.2 ml well plates (Thermo Scientific Abgene, USA). Then, 32µl nuclease-free water (Ambion®, Applied Biosystems, Thermo Fisher Scientific, USA) was added to result in a total volume of 50µl. This volume was chosen in consideration of the volume of the beads. Subsequently, 40µl (0.8x sample volume) SPRIselect® reagent magnetic bead solution was added to the samples, which were then mixed by trituration (10x) to create a homogenous solution, followed by an incubation for 5 minutes at room temperature. Next, the plate was placed on a magnetic plate (Agencourt® SPRIPlate 96R - Ring Super Magnet Plate, Beckman Coulter, UK) for 3 minutes or until the solution became clear. Keeping the plate on the magnet, supernatant (~85µl) from each sample was transferred to a new nuclease-free 1.2 ml well plate (Thermo Scientific Abgene, USA). Afterwards, 35µl (0.7x starting volume) SPRIselect® reagent was added to and mixed with the samples by trituration (10x), then incubated for 5 minutes at room temperature. The plate was then placed on the magnet and incubated for 3 minutes to attract the beads to the bottom of the wells, leaving a clear supernatant on top. Next, keeping the plate on the magnet, the supernatant was discarded carefully, without disturbing the beads. This was followed by a washing step, with the plate on the magnet, with 200µl 80% ethanol (freshly prepared), which was made up by mixing calculated volumes of absolute molecular biology grade ethanol (Fisher Scientific, UK) with nuclease-free water (Ambion®, Applied Biosystems, Thermo Fisher Scientific, USA). Plates were incubated for 30 seconds, with the ethanol then discarded. The washing step was repeated, without disturbing the bead pellet, then the wells were left to air dry for 3 minutes at room temperature to remove any residual ethanol. The plate was removed from the magnet plate, then 25µl nuclease-free water (Ambion®, Applied Biosystems, Thermo Fisher Scientific, USA**)** was added directly

onto and triturated with the bead pellet to re-suspend it, before placing the plate on the magnet plate. Next, the supernatant, which contains the purified size selected library (PSS), was transferred to fresh 0.2 ml sterilised nuclease free tube strips (BIOplastics, the Netherlands) and 2µl was taken for library quantitation by the 2100 Bioanalyzer® instrument.

2.2.4.11. Library quantitation using 2100 Bioanalyzer® instrument for Ion Template preparation.

In order to proceed with template preparation, using the Ion OneTouch™ System (Life Technologies, Paisley, UK) to prepare enriched template-positive Ion OneTouch™ 200 Ion Sphere™ Particles (ISPs; Life Technologies, Paisley, UK), the library had to be quantified and subsequently diluted. Dilution of the library ensures that the optimal input concentration of the library required for template preparation is acquired, and thus, the recommended ideal percentage (10-30%) of template-positive Ion OneTouch™ 200 Ion Sphere™ Particles (ISPs), which subsequently affects the quality of the sequencing, is achieved. Less than 10% indicates an insufficient number of template ISPs, which would prevent optimal loading density, leading to less data. On the other hand, a percentage above 30% would lead to polyclonal fragments (multi-template ISPs) on beads instead of a single population on each bead (monoclonal).

Next, barcoded and size selected libraries were pooled in equimolar amounts, to ensure equal representation of each barcoded library in the sequencing run. The dilution factor for the libraries was calculated from the molar concentration (pmol/l) of sample obtained from running the PSS on the 2100 Bioanalyzer® instrument (Figure 2.14). The recommended dilution factor led to a concentration of 26pM, which is suitable for downstream template preparation. While this concentration was recommended by the manufacturer in the 2013 version of the protocol, it has since been amended to 100pM

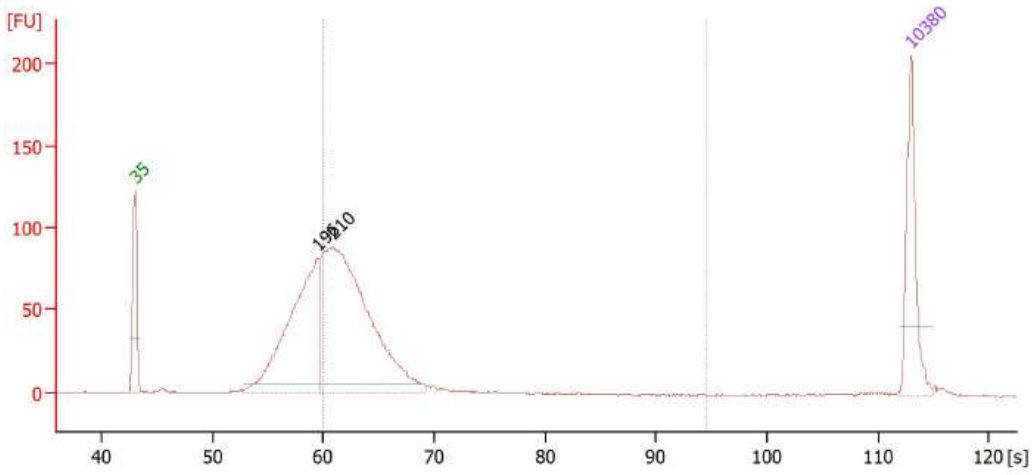
to optimize results. The manufacturer also suggested that the 2100 Bioanalyzer® instrument may produce less accurate quantitation data than, for instance, qPCR. Therefore, it was suggested that 3 serial dilutions be conducted: 1x library dilution (26pM); 0.5x dilution (13pM); and 2x dilution (52pM). Proportional dilutions were also applied to the updated concentration of 100pM. The actual recommended concentration was performed and, according to the percentage of template-positive Ion PGM™ Template OT2 200 Ion Sphere™ Particles (ISPs), the concentration was re-calculated as follows:

Dilution factor = (library concentration pM/26 or pM/ 100).

The library was diluted according the above dilution factor. For example, with a dilution factor 50, 1µl of library was added to 49µl nuclease-free water (Ambion®, Applied Biosystems, Thermo Fisher Scientific, USA). Then, to ensure equal representation of each barcoded library, equal volumes (5 µl) of each diluted library were pooled into a 1.5ml Eppendorf LoBind® tube and stored at 4^{o}C (to use within 48 hours), while the rest (undiluted libraries) were stored at -20 ^{o}C.

2.2.5 Template preparation

The process of template preparation was performed by Dr Michele Kiernan in the Systems Biology Centre at Plymouth University. For template preparation, the Ion OneTouch™ 2 System (Life Technologies, Paisley, UK) was used to prepare enriched, template-positive Ion PGM™ Template OT2 200 Ion Sphere™ Particles with clonally amplified DNA (200bp average insert libraries), by the Ion PGM Template OT2 200 Kit and used on the Ion Personal Genome Machine® (Ion PGM™) System. The Ion OneTouch™ 2 System includes: the Ion OneTouch™ 2 Instrument, and the Ion OneTouch™ ES Instrument (Life technologies, Paisley, UK).



Overall Results for sample 1 : P SS

Number of peaks found:	2	Corr. Area 1:	740.4
Noise:	0.5		

Peak table for sample 1 : P SS

Peak	Size [bp]	Conc. [pg/µl]	Molarity [pmol/l]	Observations
1	35	125.00	5,411.3	Lower Marker
2	196	333.66	2,580.5	
3	210	559.46	4,037.0	
4	10,380	75.00	10.9	Upper Marker

Figure 2.14. A purified size selected (PSS) *JK* sample assessed using the 2100 Bioanalyzer® instrument. The molarity of the two peaks were added to result in a total molarity of 6617.5 pmol/l or 6617.5 pM.

Template-positive ISPs, containing clonally amplified DNA, were prepared using the Ion PGM Template OT2 200 Kit (Life technologies, Paisley, UK), for 200 base-read libraries, with the Ion OneTouch™ 2 Instrument (which is based on emulsion PCR (emPCR)). First, the Ion OneTouch™ 2 Instrument was initialised by installing the 2 Ion OneTouch™ Recovery Tubes, the Ion OneTouch™ Recovery Router, amplification plate, disposable injector and the Ion OneTouch™ Oil and the Ion PGM™ OT2 Recovery Solution (reagents tubes). Then, the following reagents were processed for the amplification solution, by equilibrating to room temperature and vortexing: Ion PGM™ Template OT2 200 reagent Mix, Ion PGM™ Template OT2 200 PCR reagent B, Enzyme Mix and Ion PGM™ Template OT2 200 Ion Sphere™ particles (the latter two were not vortexed). Subsequently, the pooled library was diluted by mixing 6.5µl

library with 18.5µl nuclease-free water (Ambion®, Applied Biosystems, Thermo Fisher Scientific, USA) to reach a total volume of 25µl, which was then vortexed for 5 seconds and spun for 2 seconds, before being placed on ice to use within 48 hours. Next, in a 1.5-mL Eppendorf LoBind® Tube, the amplification solution was prepared by adding, in a designated order, the following: 25µl nuclease-free water; 500µl Ion PGM™ Template OT2 200 Reagent Mix; 300µl Ion PGM™ Template OT2 200 PCR Reagent B; 50 µl of Ion PGM™ Template OT2 200 Enzyme Mix; and 25 µl diluted library. The amplification solution was then vortexed vigorously for 5 seconds, followed by brief centrifugation for 2 seconds. Subsequently, 100µl Ion PGM™ Template OT2 200 Ion Sphere™ Particles (ISPs), (pre-mixed) was added to the amplification solution, which was then vortexed thoroughly for 5 seconds and dispensed into the Ion OneTouch™ 2 instrument, within 15 minutes. The entire solution (1ml) was then added to the Ion OneTouch™ 2 instrument, to which 1.5ml Ion OneTouch™ Reaction Oil was also added, before commencing the amplification reaction (which also involves centrifugation). The template-positive ISPs were recovered, within 16 hours, by carefully removing almost all of the solution from two recovery tubes, leaving 50µl recovery solution (containing the ISPs pellets). These pellets were re-suspended by trituration and together transferred into the first well of the 8-well strip provided Ion OneTouch™ ES Supplies Kit, for enrichment. However, this solution can also be stored for 3 days at 2-8 °C combined with 1 ml Ion OneTouch™ Wash Solution.

2µl amplification solution was used to assess the quality of the unenriched, template-positive ISPs, using the Ion Sphere™ Quality Control assay on the Qubit® 2.0 Fluorometer to check the percentage of template-positive ISPs. The optimum data output for template-positive ISPs is 10-30%; although, a percentage outside this range may still provide good data, as suggested by the manufacturer. The enrichment of Template-positive ISPs with Ion OneTouch™ ES involved adding reagents into the 8-

well strip well (containing 100µl unenriched template-positive ISPs in well 1) in the following order: 130µl Dynabeads® MyOne™ prepared solution; 300µl of Ion OneTouch™ Wash Solution (W); 300µl Ion OneTouch™ Wash Solution (W); 300µl Ion OneTouch™ Wash Solution (W) and 300µl Ion OneTouch™ Wash Solution (W). 300 µl freshly-prepared Melt-Off Solution was added to well 7, and wells 6 and 8 were left empty. After completing the enrichment step, ~230 µl solution containing the enriched ISPs were next used for sequencing, using the Ion PGM™ Sequencing 200 Kit v2.

2.2.6 Sequencing

The chip loading and sequencing steps were conducted by Dr Michele Kiernan who runs the Genomics Facility of the Systems Biology Centre. Templates were loaded onto the 316^{TM} chip after processing with the Ion PGM™ 200 Sequencing Kit v2 (Life technologies, Paisley, UK). Although a 314 ^{TM} chip would have sufficed for this protocol, loading onto that model of chip was unsuccessful due to the size of its wells (section 3.4.1). Samples were then sequenced using Ion Torrent PGM^{TM} (Life technologies, Paisley, UK). The sequencing protocol started by preparing the enriched, template-positive ISPs as follows: 5µl Control Ion Sphere™ Particles was added to a 0.2-ml non-polystyrene PCR tube and triturated thoroughly, followed by centrifugation for 2 minutes at 15,500×g, which is the first step of annealing the enriched, template-positive ISPs with the sequencing primers. Then, the supernatant was carefully removed without disturbing the pellet, leaving 15µl in the tube (confirmed by visually comparing to 15 µl of water in a separate tube). Next, 12µl Sequencing Primer was added, following which the solution was triturated and placed in a thermal cycler set up as follows: 95°C for 2 minutes; 37°C for 2 minutes, after which the reaction was kept at room temperature and the chip integrity was checked. The chip check involved carefully placing the chip onto the Ion PGM™ Sequencer grounding plate, or in the Ion

centrifuge adapter/rotor bucket and following the testing protocol, as instructed by the manufacturer. It was crucial to ensure the chip was not damaged, faulty or leaky and was functioning properly prior to loading the samples. Subsequently, the reaction (the enriched template-positive ISPs with annealed Sequencing Primer) was processed, via a binding reaction, by adding 3µl Ion PGM™ Sequencing 200 v2 Polymerase to the solution, which was then triturated and incubated for 5 minutes. Next, this solution (30µl) was carefully loaded into a 316 ™ chip (Life Technologies, Paisley, UK) after discarding all the liquid in the chip. The chip was then placed into the Ion PGM machine. The instrument was programmed with the necessary details for the run and with the required plugins.

2.2.7 Data analysis and bioinformatics

The raw sequence data generated from each Ion PGM™ run was transferred to the Ion Torrent server, then the Ion Torrent Suite™ converted data into different file formats, such as VCF (Variant Calling Files), FASTQ files and BAM files. In addition, a report illustrating a summary of the sequencing status, including the percentage loading onto the chip was provided by Ion Torrent Suite™. These were analysed using bioinformatics software, by aligning the resulting reads with the reference *FY* gene (NM_002036) and *JK* (NM_001146036.2), which has 11 exons and *ABO* (NM_020469.2) sequence and human genome 19 (hg19).

Alleles and variants were analysed and visualised using the following software packages and plugins: Ion Torrent Suite™ plugins (such as coverage analysis v3.6.63324, VariantCaller v3.6.59049 and FastQC v3.4.1.1), Integrative Genomics Viewer (IGV) and Ion Reporter™. In addition, the SeattleSeq annotation 137 online tool was used for data analysis (available at http://snp.gs.washington.edu/SeattleSeqAnnotation137/index.jsp). These software packages were connected to a database, such as 1000 genome

(http://www.1000genomes.org) and dbSNP (http://www.ncbi.nlm.nih.gov/SNP), in order to identify variants and SNPs.

The sequence data obtained by the Ion Torrent server had average coverage depths of 5600x for *FY*; 750x for *JK*; and 650x for *ABO*. The above software, linked to database such as the 1000 genome database (http://www.1000genomes.org), was used to align and analyse these reads with the reference sequences for the *JK* gene (NM_001146036.2), *FY* gene (NM_002036) and the *ABO* gene (NM_020469.2) from the human genome database (hg19).

2.3 *JK*: cDNA analysis of G810A (exon 8 near splice region) and A588G (exon 7 with *JK*01W*).

2.3.1 (*JK*) cDNA analysis of G810A

This analysis required the blood samples to be as fresh as possible, due to the vulnerability of RNA; therefore, the most fresh fully typed (Kidd included) blood units used here were 4 days old. The extraction of gDNA and RNA was conducted in parallel, so as not to affect RNA extraction by any possible delays. Since the first step of the experiment involved RNA, gDNA, once extracted, was stored. Sample RNA was quantified prior to synthesising cDNA, followed by amplification using specific PCR primers. These primers were designed to start (attach) within exons 8 and 9, due to an absence of introns in the messenger RNA. Finally, the cDNA was sequenced by Sanger sequencing and then analysed (Figure 2.15).

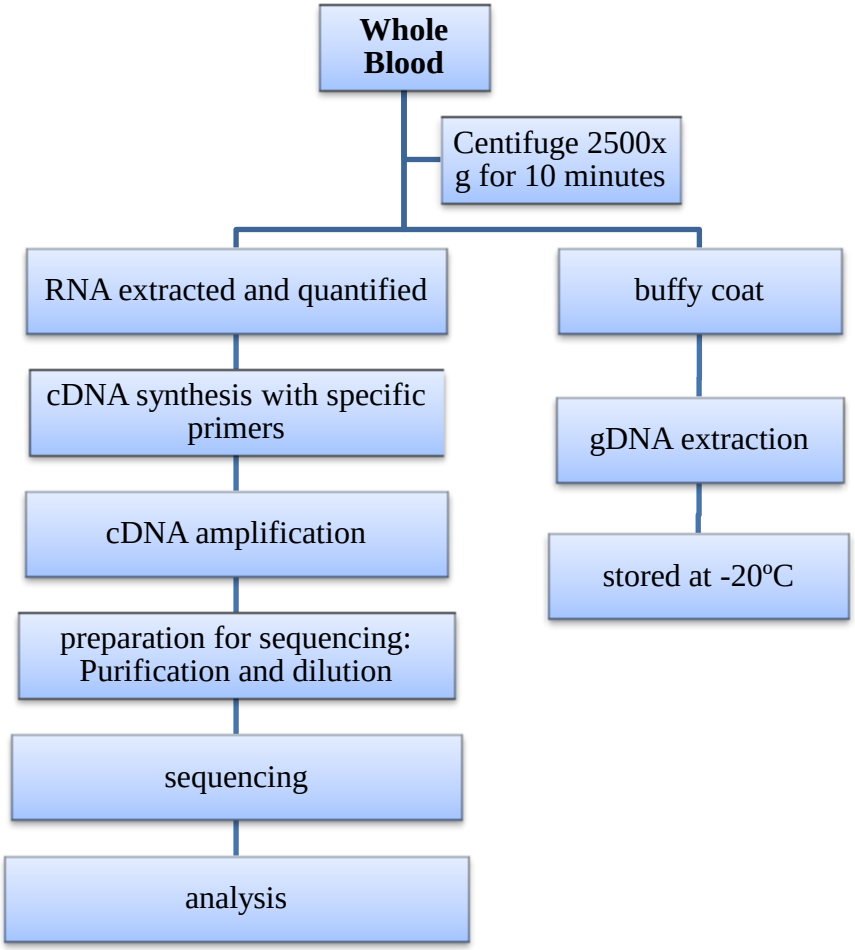


Figure 2.15. Overview of the G810A analysis

2.3.2 (*JK*) cDNA analysis of A588G

The workflow for this protocol was the same as that in section 2.3.1, until the point of amplification of the cDNA, for which different specific PCR primers were used. These primers were designed to cover *JK* exons 4-9 to include the following SNPs: G130A, A588G, G810A and G838A. These SNPs were used to analyse the association of A588G in exon 7 with *JK*01W*. Molecular cloning was then carried out in order to be able to sequence individual alleles of interest and possibly confirm the association of A588G with the *JK*01W* allele (sequenced by Sanger method; Figure 2.16).

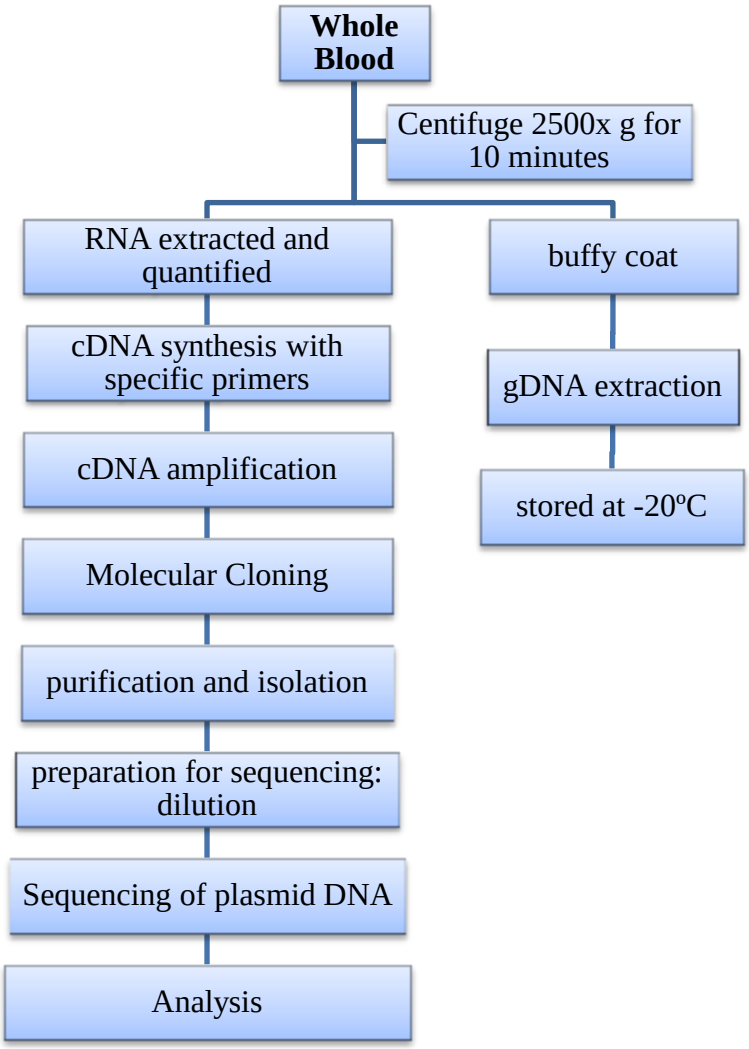


Figure 2.16. Overview of the A588G analysis

2.3.3 Primers

Several primers were designed in order to evaluate the best pair to amplify the cDNA from the required areas of the gene (exon 8 and 9; Table 2.9) and (exon 4 to 9; Table 2.10). Due to the fact that, naturally, the introns are spliced out from the RNA in the transcription process, the primers need to be designed to anneal within the exons (See chapter *JK*; figure 4.13).

Therefore, several databases were used to design and confirm the specificity of the primers, namely: Primer 3 software (http://frodo.wi.mit.edu/primer3); the NCBI Blast database (http://blast.ncbi.nlm.nih.gov/), and the UCSC Genome Bioinformatics database (https://genome.ucsc.edu). In particular, in-silico PCR (virtual PCR; https://genome.ucsc.edu) was used to confirm the specificity of primers and visualization of the amplicon.

Table 2.9. The primers designed to cover the gene region within exons 8 and 9 of *JK* for the analysis of G810A (exon 8 near splice region). Tm: melting temperature, as indicated by the NCBI BLAST database. More than 3 pairs of primers were analysed; the primers indicated here were selected as they displayed similar Tm and specificity. From these primers, FR5 (highlighted in grey) was the pair selected for the current protocol, and was used at a concentration of 500nM.

Primer	Sequence (5'-3')	Size bp	Tm(ºC)	Amplicon size with intron (bp)	Amplicon size without intron (bp)
F1	CCAGTGGGAGTTGGTCAGAT	20	59.01	477	260
R1	AGAGCCAGGAGGTGGGTTT	19	60.46		
F5	ACCAGTGGGAGTTGGTCAGA	20	60.40	454	237
R5	GTGAGCGCCATGAACATTCC	20	59.90		
F6	ATGGACAGGGGGCATTTTCC	20	60.33	440	223
R6	AAGAGCCAGGAGGTGGGTTT	20	61.07		

Table 2.10. The primers designed to cover the gene region within exons 4 to 9 of *JK* for the analysis of A588G in exon 7 with *JK*01W*.

Primer	Sequence (5'-3')	Size bp	Tm(ºC)	Amplicon size with intron (bp)	Amplicon size without intron (bp)
F1	TGGTTAGGGGTGAAAACCAG	20	57.33	9260	864
R1	CCATGAACATTCCTCCCATT	20	55.35		

2.3.4 cDNA synthesis and specific amplification

Briefly, cDNA generation is catalysed by reverse transcriptase (RT), which synthesises cDNA from messenger RNA (mRNA) via complementary base pairing of free nucleotides to single-stranded mRNA.

Using RT-PCR, synthesis of the first cDNA strand was conducted using the Superscript ® III First Synthesis System (Invitrogen™, Paisley, UK), which contains the SuperScript® III Reverse Transcriptase enzyme. To start this process, 40ng extracted RNA (~4µl, depending on the concentration of the extracted RNA) was added together with the reagents included in the kit, namely 1µl 50µM oligo(dT) and 1µl 10 mM dNTP, in 0.2ml sterile strips. DEPC-treated water was added to reach a total volume of 10µl (i.e. if 4 µl RNA was used, 4 µl DEPC-treated water was added). The solution was then briefly mixed and centrifuged, followed by incubation at 65°C for 5 minutes in a thermocycler machine and subsequent cooling on ice for at least 1 minute. Next, the cDNA synthesis mix was prepared by adding the following components (provided in the kit), in the following order: 2µl 10X RT buffer; 4µl 25mM $MgCl_2$; 2 µl 0.1 M DTT; 1µl RNaseOUT™ (40 U/µl) and 1µl SuperScript® III RT (200 U/µl). To avoid pipetting errors, the master mix was made up with 10% extra volume. Of this cDNA mix, 10µl was added to the RNA-containing mix and gently triturated, followed by brief centrifugation. The cDNA-RNA mix was then incubated in the thermocycler machine at 50°C for 50 minutes and the reaction was terminated at 85°C for 5 minutes, before the tubes were placed on ice. Afterwards, the tubes were spun and 1µl of RNase H, which removes the remaining RNA, was added to the tubes, following which they were incubated at 37°C for 20 minutes in the thermocycler. A working aliquot of 6µl first-strand cDNA was used for subsequent experiments, while the rest was stored at – 20°C.

2.3.5 Amplification of target cDNA

The synthesised first-strand cDNA was next amplified by PCR, where *JK* exon 8-9- and *JK* exon 4-9-specific primers were utilised to enable amplification of that region of the gene and amplicon quantification, thereby facilitating downstream analysis. In this case, successful amplification would indicate successful synthesis of cDNA from RNA. 50µl amplification reaction mix (containing primers flanking *JK* exons 8-9 for the analysis of the G810A on exon 8, near splice region) contained the following: 4-7µl cDNA; Q5® Hot Start High-Fidelity 1X Master Mix (New England BioLabs $_{inc}$, UK); 500nM primers; and a volume of nuclease-free water required to bring the total volume to 50µl (depending on the volume of cDNA added). Prior to amplification, several optimisation tests were carried out (data not shown), which included changes in temperature, times and other thermocycling conditions. Thermocycling was carried out using a Veriti Thermal Cycler (Life Technologies, Paisley, UK), under conditions indicated in Table 2.11. Subsequently, amplicons were visualised by agarose gel electrophoresis, using 1.5% Hi-Res Standard agarose gel (Geneflow Limited, Staffordshire, UK). 10µl amplicons was mixed with 2µl DNA gel loading buffer and loaded into gels, following which electrophoresis was carried out at 90V for 1 hour and bands were visualized and compared against TriDye^{TM} 100 bp (New England BioLabs® inc., UK). Amplicons then were then sequenced using the Sanger method (section 2.3.6).

Table 2.11. Thermocycling conditions for cDNA amplification, specific for the analysis of the *JK* (G810A on exon 8 near splice region)

Step	Temperature	Time	Cycles
Initial denaturation	98°C	30 seconds	1
Denaturation	98°C	5 seconds	35
Annealing	64°C	20 seconds	
Extension	72°C	30 seconds	
Final extension	72°C	2 minutes	1
Holding	4°C	∞	-

Amplification of target cDNA that covers *JK* exon 4-9, including SNPs G130A, A588G, G810A and G838A (for the analysis of the association of A588G in exon 7 with *JK*01W*) was different for this particular amplicon, as molecular cloning was required. The 50µl amplification reaction mix included 80ng cDNA and BioMix™ 2x reaction mix (BioLine, UK), with a primer concentration of 500nM (Table 2.10), mixed in 0.2 ml tube strips. The BioMix™ 2x reaction mix contains a stable Taq DNA that leaves the nucleotide A overhang on the PCR product, which is critical for the cloning process. Thermocycling was carried out using a Veriti Thermal Cycler (Life Technologies, Paisley, UK), with the optimised conditions indicated in Table 2.12. The final extension was prolonged to 20 minutes to ensure that all cDNA amplicons were of full length, with nucleotide A overhang at the 3' end. The amplicons were then visualised by agarose gel electrophoresis, using 1% Hi-Res Standard agarose gel (Geneflow Limited, Staffordshire, UK), with electrophoresis carried out at 90V for 1 hour. Amplicons showing single bands were then prepared for molecular cloning and Sanger sequencing.

Table 2.12. Optimised thermocycling conditions for cDNA amplification, specific for analysis of *JK* A588G in exon 7. The final extension was prolonged by 20 minutes to ensure full length cDNA with nucleotide A overhang at the 3' end.

Step	Temperature	Time	Cycles
Initial denaturation	94^{o}C	30 seconds	1
Denaturation	94^{o}C	5 seconds	35
Annealing	60^{o}C	20 seconds	
Extension	72^{o}C	30 seconds	
Final extension	72^{o}C	20 minutes	1
Holding	4^{o}C	∞	-

2.3.6 cDNA Sequencing (Sanger) of amplicon covering G810A

The cDNA amplicons for the analysis of the G810A were purified and diluted according to the manufacturer's instructions for the SmartSeq Kit (Eurofins Genomics), whose name subsequently changed to Mix2Seq kit. http://www.eurofinsgenomics.eu. This kit contains pre-labelled, barcoded tubes with secure lid sealing, which are recognised by Eurofins company, in which the purified template (here was cDNA amplicon) and specific primer were premixed and loaded before sending them for sequencing. The amplicons were purified by SPRIselect® reagent kit with 80% of freshly prepared molecular biology grade ethanol (Fisher Scientific, UK). In a plate, the remaining volume of cDNA (40 µl) were mixed thoroughly with 1.8X SPRIselect® magnetic beads (72 µl) by pipetting 10 times and incubated at room temperature for 10 minutes. Next, the plate was placed on the magnetic plate (Agencourt® SPRIPlate 96R - Ring Super Magnet Plate, Beckman Coulter, UK) for 3 minutes with which the Amplicon-carrying beads attached to the wall of the wells leaving a clear solution supernatant on top, which was then carefully discarded; however the supernatant of the very first experiment of

the cDNA purification was kept as due to small size (237bp), there was a concern that there may be a failure in the beads attracting the small amplicons. The supernatant of the first cDNA amplicon purification experiment was tested and revealed that the amplicons were successfully attracted by beads as showing negative in the gel of the supernatant (data not shown). Subsequently, without removing the plate from the magnet, a 2 times wash was conducted by the addition of 30µl of the 80% ethanol, then incubated for 30 seconds and discarded. After that, an air-dry for 3 minutes was performed before removing the plate from the magnet and purified amplicons were eluted by 30µl of nuclease-free water (Ambion®, Applied Biosystems, Thermo Fisher Scientific, USA) as explained in several sections. Afterwards, those purified cDNA samples were quantified by Qubit® dsDNA Broad range (BR) assay Kit (Invitrogen™, Paisley, UK), in order to calculate the dilutions to meet the requirements of the SmartSeq Kit or Mix2Seq kit (Eurofins Genomics). For optimum Sanger sequencing results using SmartSeq Kit or Mix2Seq kit (Eurofins Genomics), 15 µl with concentration of (1ng/ µl) of purified PCR product (cDNA amplicons in this case) was required to mix with 2 µl of 10pmol/µl each forward and reverse primer. Therefore, 2 diluted aliquots, as two sequencing reactions are needed so each sample would be sequenced on both strands with forward and reverse primer. Each aliquot was prepared with volume of 20 µl (with 1ng/ µl), to avoid pipetting error, by the following equation: volume to be taken from the purified cDNA amplicon = 1(ng)x 20(µl)/concentration of the cDNA amplicon (from the Qubit®) and make up to 20 µl with nuclease free water. On the other hand, 10pmol/ µl aliquot with volume made up to 50 µl with nuclease free water of each forward and reverse primer (specific primers for G810A analysis). In two 1.5 ml Eppendorf tubes for each sample, 15 µl of the (1ng/ µl) purified cDNA amplicons were mixed with 2 µl of (10pmol/ µl) of forward or reverse primers in indicated tubes ensuring the volume of this mix to at least 17µl. Then, these premixed

solutions were transferred and loaded into the barcoded tubes provided in the Mix2Seq kit (Eurofins Genomics) and sealed before sending them to the designated company (Eurofins Genomics) address to be Sanger sequenced.

2.3.7 Molecular cloning

The molecular cloning process utilised TOPO® TA Cloning® Kit for sequencing (Invitrogen™, Paisley, UK) with pCR™ 4-TOPO® vector (TOPO®vector)- and Top10 One Shot® Chemically-competent cells, and was conducted near a Bunsen Burner. First, the TOPO® Cloning reaction (6µl) was performed by adding the following reactants into 0.2 ml nuclease-free tube strips, in the indicated order: 4µl freshly prepared PCR product (Taq-amplified cDNA); 1µl Salt solution (provided in the kit); and 1µl TOPO®vector, which enables ligation of the PCR product (insert) to plasmid (recombinant) DNA. This reaction was then gently mixed, followed by incubation for 15 minutes at room temperature and placement on ice for the chemical transformation process.

The mixed cloning reactants were transformed into Top10 One Shot® Chemically competent cells as follows: 2µl TOPO® Cloning reaction was carefully added into a vial of One Shot® chemically competent *E. coli* and then mixed by gently rotating, followed by incubation on ice for 20 minutes. Subsequently, the vial was subjected to a heat-shock at 42°C for 30 seconds (without shaking) and immediately replaced on ice. Next, 250µl S.O.C. medium (provided in the kit and equilibrated to room temperature) was carefully added to the vial, which was then placed horizontally into a shaking (200 rpm) incubator at 37°C for 1 hour. Then, in 5 separate 1.5ml sterile Eppendorf tubes, 50µl of this transformation solution was mixed with 20µl S.O.C. medium to homogenously spread the transformation solution over 5 pre-warmed selective plates (LB plates containing 50µg/ml Kanamycin). Plates were incubated overnight at 37°C. The reason for using 5 separate plates was to use up all the transformation solution (250µl) and

thereby culturing all possible colonies, so as not to miss a colony containing the needed insert (*JK*01W* allele). Subsequently, 30 colonies were randomly selected from the 5 plates and sub-cultured, by inoculating a minute amount (using the tip of a sterile loop) into a graded selective plate, followed by overnight incubation (growth) at 37°C and subsequent storage at 2-8°C.

Meanwhile, the majority of the 30 colonies were cultured overnight in individual tubes containing 5ml medium (LB medium contacting 50µg/ml Kanamycin) at 37°C in a shaking (80rpm) incubator. Then, the cultured colonies with recombinant DNA (Plasmid DNA with the insert) were isolated and purified using the PureLink®Quick Plasmid Miniprep Kit (Invitrogen^{TM}, Paisley, UK).

2.3.7.1 Isolation and purification of plasmid DNA

Purification of plasmid DNA allowed analysis for positive insertion of the cDNA and was required prior to sequencing. During this procedure, 4.5ml LB-culture inoculated (overnight) broth, was centrifuged at 4500 rpm for 10 minutes at room temperature. 500µl of LB-culture broth was stored together with 15% (v/v) sterile glycerol at -80°C for future use. Following centrifugation, the supernatant was removed gently, without disturbing the pellet, and discarded. The pellet was re-suspended by adding 250µl Resuspension Buffer with RNase A (R3; provided in the kit) and solution was homogenised by trituration. Cells were then lysed by adding 250µl Lysis Buffer L7 (provided in the kit) to the tube, which was then gently mixed by inversion until homogenous and incubated at room temperature for 5 minutes. Next, the solution was precipitated by adding 350µl Precipitation Buffer N4 (provided in the kit) and immediately mixed by inversion, until a turbid white homogenous solution formed, which was then centrifuged at >12000xg for 10 minutes. The supernatant was loaded onto a (provided) spin column (placed in 2ml wash tubes) for the binding process and spun for 1 minute at 12000xg, after which the flow-through was discarded and the

column replaced into the washing tube. The column was then washed using 500µl Wash buffer (W10; with ethanol), incubated for 1 minute at room temperature, then spun again for 1 minute at 12000xg, before discarding the flow-through. Another wash was performed as above, this time 700µl Wash buffer (W9; with ethanol), following which the column was centrifuged again, without the addition of any buffers, at 12000xg for 1 minute and flow-through discarded. Columns were then placed in fresh 1.5 ml recovery tubes (provided in kit). Subsequently, 75µl TE buffer (preheated to 65-70°C) was added to the centre of each column for elution, and then incubated for 1 minute at room temperature. Columns were then centrifuged at 12000xg for 2 minutes and discarded, while the recovery tubes (containing purified plasmid DNA) were equally aliquoted and stored at - 20°C. Purified plasmid DNA was quantified using the Qubit® 2.0 Fluorometer and the Qubit® dsDNA Broad range (BR) assay Kit (Life Technologies, Paisley, UK), prior to screening for plasmid DNA with positive insertions, which was conducted via restriction enzyme (RE) analysis using *EcoRI* (Promega, UK).

2.3.7.2 Restriction enzyme analysis to screen plasmid DNA for insert

The restriction digest solution contained the following reagents, added to 0.2 ml PCR tube strips in the indicated order: 15.3µl nuclease-free water; 2µl RE 10X buffer; 0.2µl Acetylated BSA (10µg/µl); 2µl purified plasmid DNA (mixed by trituration); and 0.5µl *EcoRI* (10U/µl) to result in a 20 µl reaction mix, which was subsequently triturated and incubated 37°C for 1 hour in a thermocycler. To visualise the plasmid DNA with the insert, the total restriction digest mix was loaded into a 1.5% Hi-Res Standard Agarose gel (Geneflow Limited, Staffordshire, UK), mixed with 4µl DNA gel loading buffer. In addition, 2µl uncut purified plasmid DNA was mixed with 3µl DNA gel loading buffer, which was used as control. Gel electrophoresis was carried out 90V for 1 hour, following which bands of the desired size (indicating plasmid DNA samples with positive insertions) were sequenced using the Sanger method. Plasmid DNA with

positive insertions was diluted as to be used for the SmartSeq Kit (Eurofins Genomics, Ebersberg, Germany), according to the manufacturer's instructions. For Sanger sequencing of plasmid DNA, aliquots containing DNA concentrations of 80ng/µl were prepared as previously described in 2.3.6; 15µl was loaded into 2 separate barcoded tubes with 2µl (10pmol/µl) primers (T3**;** 5´-ATTAACCCTCACTAAAGGGA-3´) and (T7; 5´-TAATACGACTCACTATAGGG-3´) provided in the TOPO® TA Cloning® Kit for Sequencing (Invitrogen™, Paisley, UK), prior to sequencing.

2.4 Sanger sequencing for *ABO* gene

In the first NGS *ABO* run, various SNPs from exons 3, 4 and 6 were investigated and by Sanger sequencing to confirm the NGS data validity. Primers were designed to amplify around SNPs in exon 3 (103G/A, 106G/A), exon 4 (188G/A, 189C/T) and exon 6 (297A/G). The primers used for amplification of these SNPs are listed in Table 2.13; these primers were used together with the Q5® Hot Start High-Fidelity 2X Master Mix (New England BioLabs $_{inc}$, UK), with thermocycling conditions described in Table 2.14.

Table 2.13 Primers designed and used to confirm number of *ABO* SNPs from NGS data.

Primer	Sequence (5'-3')	Size bp	Tm (ºC)	Amplicon size with intron (bp)	Exon
F1	CAGCTTCACCGGGAACTCTT	20	59.4	210	3
R1	ATGGATGCTCCACCTGCTCT	20	59.4		
F2	CACAGTGATGGTTGTTCTGTGA	22	58.4	368	4
R2	GGCCAAGTCAGGGTAGAGACT	21	61.8		
F3	AGAAGCTGAGTGGAGTTTCCAG	22	60.3	318	6
R3	TGAACACAAGGAGAGACCTCAAT	23	58.9		

Table 2.14 Thermocycling conditions for amplification of the primers in Table 2.13 used for the confirmation step.

Step	Temperature	Time	Cycles
Initial denaturation	98^{o}C	30 seconds	1
Denaturation	98^{o}C	5 seconds	35
Annealing	62^{o}C	20 seconds	
Extension	72^{o}C	30 seconds	
Final extension	72^{o}C	2 minutes	1
Holding	4^{o}C	∞	-

2.5 Sanger sequence analysis

Sequences of forward and reverse reactions were visualised and analysed using the software packages: Sequence Scanner software 2 (Seq scanner 2, Applied Biosystems, Paisley, UK), and Integrative Genomics Viewer (https://www.broadinstitute.org). These software packages were used for aligning and matching forward and reverse sequences with the reference sequences, providing an overview of the sequence variations, visualising SNPs, and assessing sequence quality.

Chapter 3

Genotyping of the FY Blood Group by Next Generation Sequencing

3.1 Introduction

The Duffy blood group system (FY/ISBT 006) is represented by a single gene (*FY* or *ACKR1*) (see section 1.6.1). Initially, this gene was described to have one exon (encoding 338 amino acids)(Chaudhuri et al., 1993); however, subsequently, the gene was shown to have two exons, encoding for 336 amino acids – assigned as the major (predominant) transcript (section 1.6.1) (Iwamoto et al., 1996a). The major transcript has been used here as the reference sequence, including the numbering of nucleotides and amino acids. The FY blood group system is considered to be clinically significant, as the corresponding alloantibodies have been reported to be involved in haemolytic transfusion reactions and haemolytic disease of the foetus or newborn (HDFN), in cases of incompatible blood transfusion or pregnancy (Daniels, 2013). The main Fy polymorphic antigens ($Fy^{a/b}$) are thought to be directly encoded by two main co-dominant alleles (*FY*A* and *FY*B*) which differ by a single nucleotide polymorphism (SNP) (125G>A) leading to the amino acid change, Gly42Asp. However, there are, in fact, numerous *FY* alleles; according to the NCBI Blood Group Antigen Gene Mutation Database (dbSNP; www.ncbi.nlm.nih.com), there are 16 *FY* alleles, a number of which impair or demolish (null phenotype) the expression of Fy antigens. Examples of these are *FY*X* (*FY*02M.01*), associated with the SNPs 265C>T and 298G>A, which weakens the expression of Fy^b antigen (Olsson et al., 1998), and the *FY*Null* associated with SNP (-67T>C) in the GATA promoter region, found commonly among African Americans (Reid et al., 2012, Tournamille et al., 1995). The number of *FY* alleles

continues to increase, many of which may have not been included in the NCBI database yet – especially low frequency alleles (Westoff et al., 2014). These new alleles arise from different mutation mechanisms, such as SNPs and deletions, which may alter the amino acid sequence or lead to a premature stop codon, thereby affecting antigenicity. Accordingly, it is suggested that providing extended genotyping of clinically significant blood group systems is suggested to improve the safety of blood transfusions and minimise alloimmunisation, especially for transfusion-dependant individuals, such as those with sickle cell disease (SCD) (Avent, 2009). Blood group genotyping (BGG) has impacted greatly on the management of alloimmunisation, by which better interpretation of the phenotype is provided (Castilho, 2007). High throughput genotyping is suggested to be vital in addressing the demand to screen for a large number of alleles (Anstee, 2009), especially those that may be rare. However, despite their high throughput capabilities, all current commercially available platforms, such as microarray-based platforms, are dependent on previous knowledge of the molecular basis of the alleles and, consequently, novel or rare mutations and alleles, which may affect the antigenicity, might be missed due to the absence of corresponding probes in the array. As a result, the genotyping system requires constant updating. On the other hand, Next Generation Sequencing (NGS) technology, which is sequence-based genotyping, circumvents the requirement of previous knowledge of, for example, SNPs, thereby allowing a discovery mode. This provides a key opportunity to thoroughly explore all existing and novel alleles, so as to provide comprehensive interpretation of the phenotype from the genotype (McBean et al., 2014, Tilley and Grimsley, 2014, Avent et al., 2015).

3.2 Aim of the study

The aim of this project was to use the powerful NGS-based genotyping platform to establish an optimised and reliable protocol for comprehensive genotyping of the entire gene of the blood group *FY* plus the flanking regions. Here, the Ion Torrent PGM™ (the available platform at Plymouth University, was used with long-range polymerase chain reaction (LR-PCR) (single amplicon). Considering the NGS capability of discovering novel polymorphisms, in addition to known ones across the gene, extensive genotyping of all existing polymorphisms in exons, introns and regulatory regions will be provided, enabling comprehensive interpretation of associated phenotypes. In addition, intronic SNPs will be analysed and catalogued in order to investigate their association with alleles, such as weak or null alleles. Accordingly, a refined illustration of the molecular basis of the *FY* alleles reference sequences will be provided, in addition to the discovery of new alleles. Moreover, this NGS genotyping and analysis approach will be evaluated, using *FY*, for its application to other blood groups in this project (*JK* and *ABO*).

3.3 Results

3.3.1 Long-range PCR of the *FY* gene

For complete *FY* genotyping by NGS, gDNA of 53 blood samples was sequenced. The full serology information of Fy^a and Fy^b phenotype was provided for 43 donor samples (Table 3.1). Using primer 3 software and the NCBI database (see section 2.2.4.1), a primer pair was designed to produce a single long amplicon (4784bp) to target the entire *FY* gene plus flanking regions (up- and downstream; Figure 3.1). For the LR-PCR amplification reaction, the forward 5'-3' GTGTGAGTGAGTGAGAGGCAGAG and reverse 5'-3' GCCAGAGAGGAGACAGAAGACAG primers (see Table 2.1) were mixed with the gDNA and 2x LongAmp® Hot Start Taq Master Mix (New England BioLabs Inc., Herts, UK), which consists of a blend of Hot Start Taq and Deep VentR® DNA Polymerases (see section 2.2.4.2). As a result, the fidelity of this master mix is two-fold greater than that of *Taq* DNA polymerase alone, according to the manufacturer's instructions. *FY* PCR amplicons were loaded onto a 1% agarose gel and electrophoresed at 85 V for 1 hour (section 2.2.4.3); amplicons were detected as a single ~5kbp band (Figure 3.2). Then, the amplicons were purified by a magnetic bead technique (see section 2.2.4.4) to ensure the removal of primer dimers and free nucleotides that may interfere with downstream processing. Subsequently, purified amplicons were quantified using the Qubit® 2.0 Fluorometer with Broad range (BR) assay Kit (Invitrogen™, Paisley, UK; section 2.2.4.4), which is crucial in order to prepare the 100 ng concentration of amplicons (Section 2.2.4.5) required for the fragmentation process. As a single amplicon was utilised to cover the *FY* gene, aliquots of 20ng/µl were prepared, from which 5µl were used for the fragmentation process.

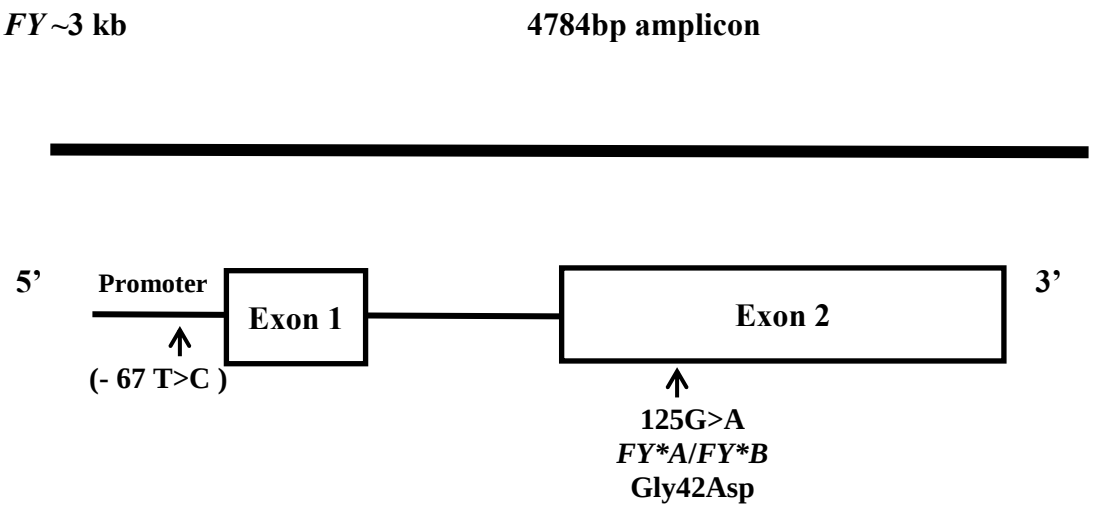


Figure 3.1. Single long amplicon (generated by LR-PCR) covers the entire *FY* gene plus flanking regions (including the promoter region).

FY gene is (2488bp according to NCBI GenBank and found as over 1.5kb in (Reid et al., 2012) with 2 exons shown as boxes while the intron is represented by a line. The main SNP that differs *FY*A*/*FY*B* is shown in exon 2, while the SNP suggested to be responsible for the null *FY* allele (*FY*02N.01*) is also shown in the promoter region.

Table 3.1. The serology information of the 53 blood samples provided from National Health Service Blood and Transplant (NHSBT; Filton, Bristol UK). * represents the sample ID provided. Note, the serology of a number of samples displays only the status of the Fy^a antigen, with the status of the Fy^b antigen being represented by ?**. ND, serology not defined. The FY008 number, which is the ISBT number assigned for the FY blood group system (Daniels, 2013, Reid et al., 2012), is used here to name the samples accordingly to the *FY* NGS genotyping experiment.

Sample number	Sample ID*	Phenotype	Sample number	Sample ID*	Phenotype	Sample number	Sample ID*	Phenotype
FY008.01	470D	ND	**FY008.21**	4648	ND	**FY008.41**	4576	Fy (a+b?**)
FY008.02	9	Fy (a+b-)	**FY008.22**	12	Fy (a-b+)	**FY008.42**	4208	Fy (a+b+)
FY008.03	25	Fy (a+b-)	**FY008.23**	15	Fy (a-b+)	**FY008.43**	581Z	ND
FY008.04	35	Fy (a+b-)	**FY008.24**	17	Fy (a-b+)	**FY008.44**	469B	ND
FY008.05	36	Fy (a+b-)	**FY008.25**	19	Fy (a-b+)	**FY008.45**	5800	ND
FY008.06	45	Fy (a+b-)	**FY008.26**	76	Fy (a-b+)	**FY008.46**	470P	Fy (a+b?**)
FY008.07	51	Fy (a+b-)	**FY008.27**	74	Fy (a-b+)	**FY008.47**	580X	ND
FY008.08	11	Fy (a+b-)	**FY008.28**	97	Fy (a-b+)	**FY008.48**	3	Fy (a+b+)
FY008.09	40	Fy (a+b-)	**FY008.29**	99	Fy (a-b+)	**FY008.49**	4	Fy (a+b+)
FY008.10	50	Fy (a+b-)	**FY008.30**	101	Fy (a-b+)	**FY008.50**	5	Fy (a+b+)
FY008.11	54	Fy (a+b-)	**FY008.31**	104	Fy (a-b+)	**FY008.51**	6	Fy (a+b+)
FY008.12	55	Fy (a+b-)	**FY008.32**	111	Fy (a-b+)	**FY008.52**	14	Fy (a+b+)
FY008.13	59	Fy (a+b-)	**FY008.33**	73	Fy (a-b+)	**FY008.53**	22	Fy (a+b+)
FY008.14	60	Fy (a+b-)	**FY008.34**	1	Fy (a-b-)			
FY008.15	71	Fy (a+b-)	**FY008.35**	2	Fy (a-b-)			
FY008.16	69	Fy (a+b-)	**FY008.36**	10	Fy (a-b-)			
FY008.17	75	Fy (a-b+)	**FY008.37**	26	Fy (a-b-)			
FY008.18	437R	Fy (a-b+)	**FY008.38**	28	Fy (a-b-)			
FY008.19	459F	ND	**FY008.39**	29	Fy (a-b-)			
FY008.20	453R	ND	**FY008.40**	31	Fy (a-b-)			

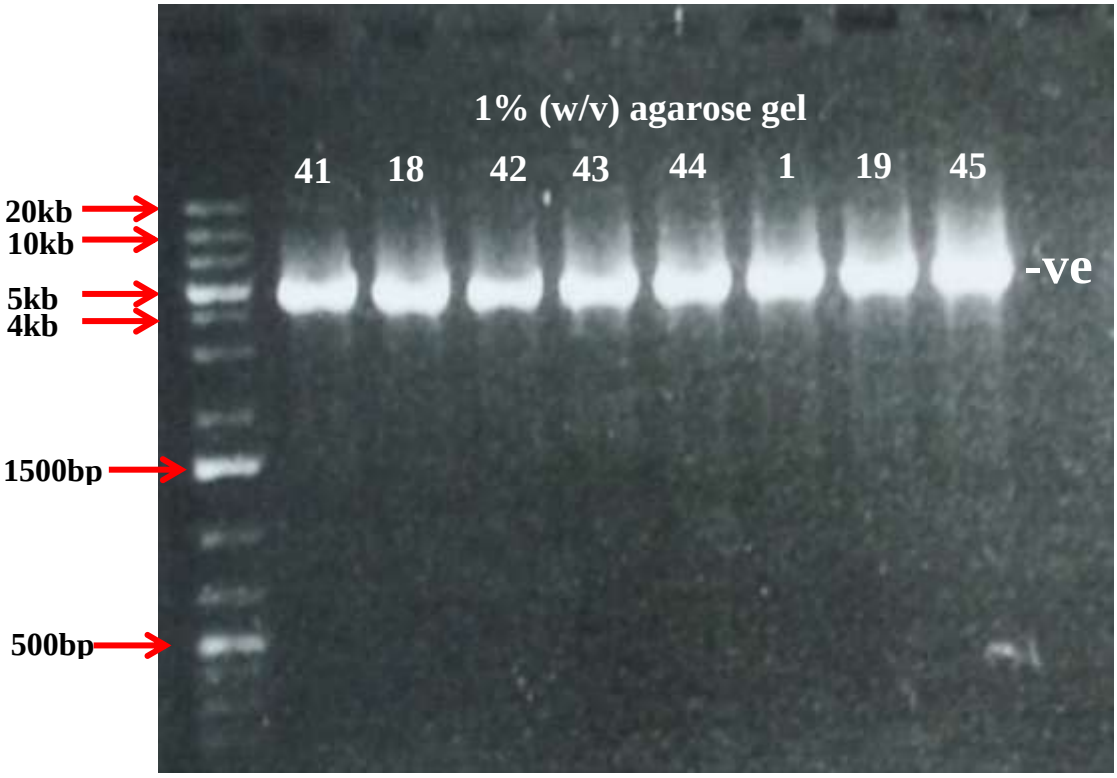


Figure 3.2. The *FY* amplicons produced by LR-PCR.

The whole *FY* gene was covered by a single amplicon of ~5kbp size (4784bp). In this example, the amplicons of 8 samples are shown. The last column (-ve) represents a negative control, which is a master mix without gDNA added. Samples were loaded on a 1% (w/v) agarose gel and electrophoresed at 85 V for 1 hour. The GeneRuler^{TM} 1Kb Plus DNA ladder (Thermo Fisher Scientific) was used as a marker of DNA size. The number of samples is shown in short form representing that after the FY008.number.

3.3.2 Next Generation Sequencing of the *FY* gene

3.3.2.1 *FY* amplicon library fragmentation (purified fragmented library)

Following amplicon purification and quantitation, NGS libraries were fragmented using the Ion Xpress™ Plus Fragment Library Kit (section 2.2.4.6). The incubation length of the enzymatic fragmentation reaction was 5 minutes, which was selected as this allows a wide fragment size distribution with sufficient yield (peak) around 200-300bp – the recommended target median fragment size by the manufacturer (Figure 3.3: an example of a *FY* fragmented-purified amplicon). Subsequently, samples were purified by magnetic beads (Agencourt® AMPure® XP beads Reagent or SPRIselect® reagent kit (Beckman Coulter, UK). Fragment size distribution of samples was assessed by the Agilent® 2100 Bioanalyzer and Agilent High Sensitivity DNA Kit (Agilent Technologies UK Limited) (Figure 3.3).

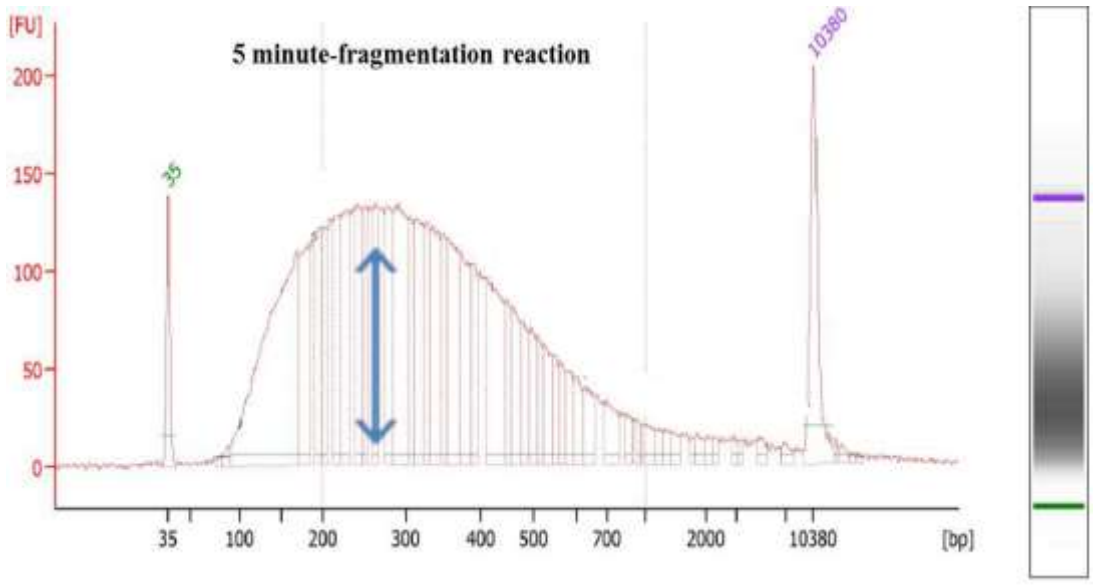


Figure 3.3. An electropherogram of a fragmented-purified *FY* DNA library.

This is an example of a *FY* amplicon fragmented using the Ion Xpress™ Plus Fragment Library Kit for 5 minutes. The arrow displays the peak around 200-300 bp. The green line (35bp) is the lower marker and the purple (10380bp) is the upper marker. Results shown were obtained using the Bioanalyzer® instrument with Agilent High Sensitivity DNA Kit (Agilent Technologies UK Limited).

3.3.2.2 Ligation of barcoded adapters (*FY* purified-ligated library)

The ligation of barcoded adapters to the fragmented samples is critical to allow the identification and distinguishing of samples after pooling. The samples were ligated to barcoded adapters (P1 and Ion Xpress™Barcode X adapter), from the Ion Xpress™ Barcode adapters Kit (Life Technologies, Paisley, UK), using DNA ligase. Subsequently, ligated libraries were purified and assessed by Agilent® 2100 Bioanalyzer and Agilent High Sensitivity DNA Kit (Agilent Technologies UK Limited). This analysis was optional and estimates the success of the adapter ligation in a form of an increase shift of size distribution by 80bp (see section 2.2.4.9). Figure 3.4 illustrates the shift in the size distribution between one *FY* fragmented library and then a *FY* fragmented-ligated library.

3.3.2.3 Size selection

The size selection process is applied to the purified ligated library to produce the length size that provides the best reads possible when processed by the Ion PGM™ Template OT2 200 kit and run on the Ion PGM™ 200 Sequencing Kit. In the process of sequencing 53 samples, three experiments were conducted. 12 samples were sequenced in the first experiment, 17 in the second and 24 samples in the third sequencing experiment. The first 12 samples underwent the majority of the library construction optimisations: fragmentation for 9 minutes; size selection by the Pippin Prep™ instrument (Sage Science, Inc., Beverly, USA) and Pippin Prep™ Kit 2010 with Ethidium Bromide cassettes; and purification by magnetic beads. This was set to target the optimum target peak (330bp), which provides the highest number of reads. However, as the fragmentation length (9 minutes) leads to a lower concentration around 200-300bp, this size selection attempt resulted in a low concentration, around 330bp (Figure 3.5), which then required an amplification step (see section 2.2.4.10).

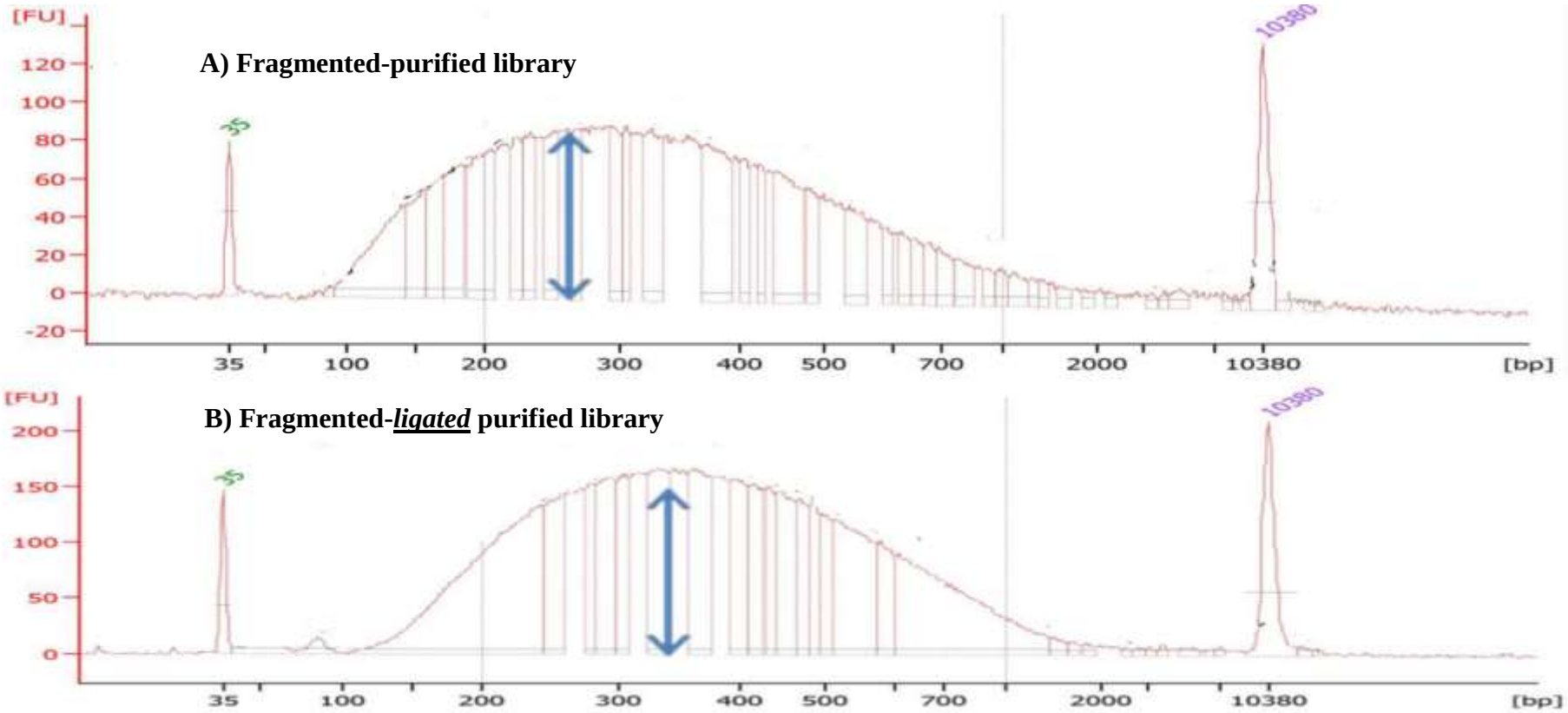
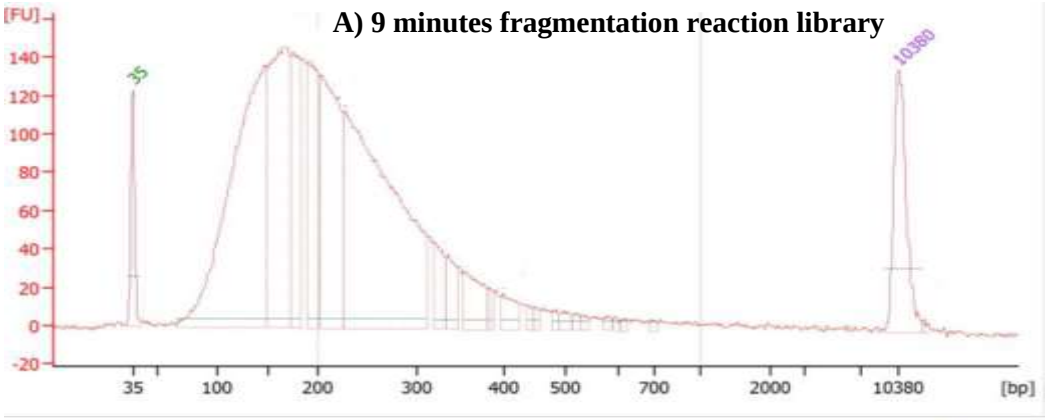
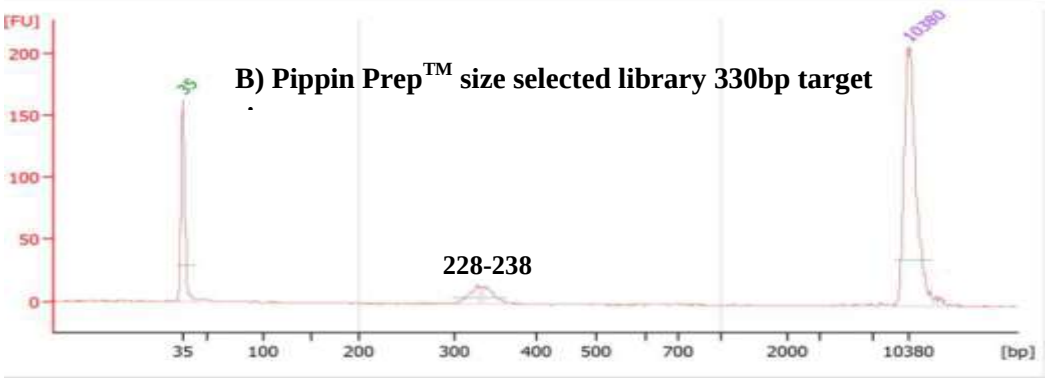


Figure 3.4. Two electropherograms of the same *FY* amplicon library.

A) Fragmented-purified library with the peak around 200-300bp
B) Fragmented-ligated purified library with the peak shifted to the right (due to adapter ligation, evident as an increase in size by ~ 80bp.

The green line (35bp) is the lower marker and the purple line (10380bp) is the upper marker. Results shown were obtained using the Bioanalyzer® instrument.





Peak		Size [bp]	Conc. [pg/µl]	Molarity [pmol/l]	Observations
1	◀	35	125.00	5,411.3	Lower Marker
2		328	11.12	51.3	
3		338	10.67	47.8	
4	▶	10,380	75.00	10.9	Upper Marker

Figure 3.5. Two electropherograms of the same *FY* amplicon library, fragmented for 9 minutes and size selected by Pippin Prep™. (Both purified)

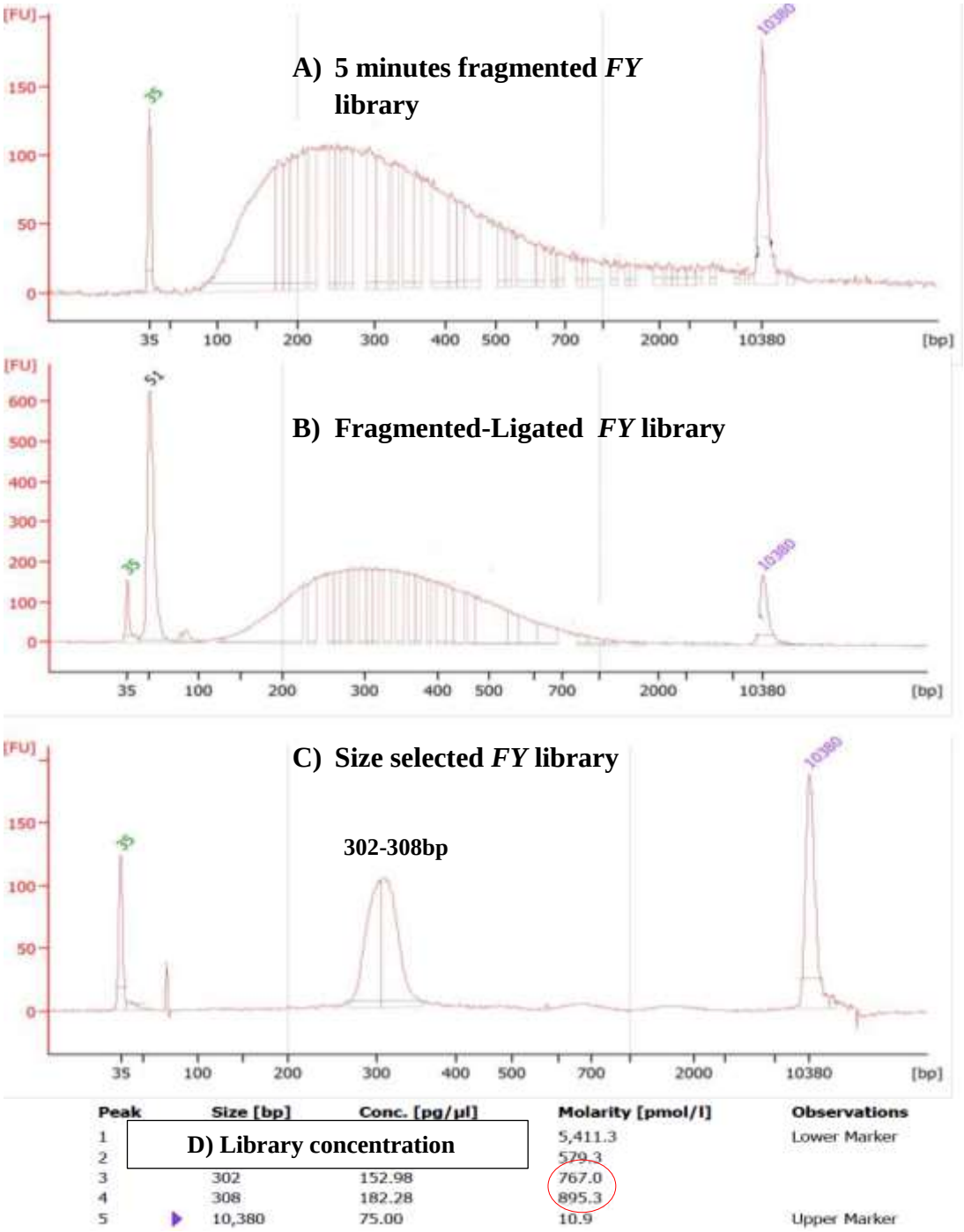
A) FY library was fragmented for 9 minutes, which clearly provides median fragment size 100-200bp.

B) Same library was size selected (by Pippin Prep™ instrument) with the required target of 330bp, which clearly provides low concentration due to low DNA distribution around this area. Note the total library concentration here is (991 pmol/l) as the actual concentration 99.1 was multiplied by 10 due to diluted library.

The green (35bp) represents the lower marker while (purple 10380) is the upper marker. Results shown were obtained using the Bioanalyzer® instrument.

Accordingly, the next set of 17 samples were fragmented for 5 minutes, in order to attain a sufficient concentration around 200-300bp, and size-selected by Pippin Prep™ instrument after ligation of the adapters. Figure 3.6 illustrates 3 electropherograms of the fragmentation, ligation and size-selection of one of the 17 *FY* libraries. The remaining 24 *FY* libraries were size-selected using SPRIselect® reagent magnetic beads (see section 2.2.4.10) to provide a peak around 200 bp, which is compatible with the Ion PGM™ Template OT2 200 kit. Figure 3.7 shows a purified-ligated *FY* library that has been size-selected by SPRIselect® reagent magnetic beads.

The size selection step of the amplicon library preparation, the concentrations of which were determined by the Bioanalyzer® instrument (see section 2.2.4.11) that were critical for template preparation (section 2.2.5), was followed by sequencing using Ion torrent PGM™ (Section 2.2.6).



Peak	Size [bp]	Conc. [pg/µl]	Molarity [pmol/l]	Observations
1			5,411.3	Lower Marker
2			579.3	
3	302	152.98	767.0	
4	308	182.28	895.3	
5	10,380	75.00	10.9	Upper Marker

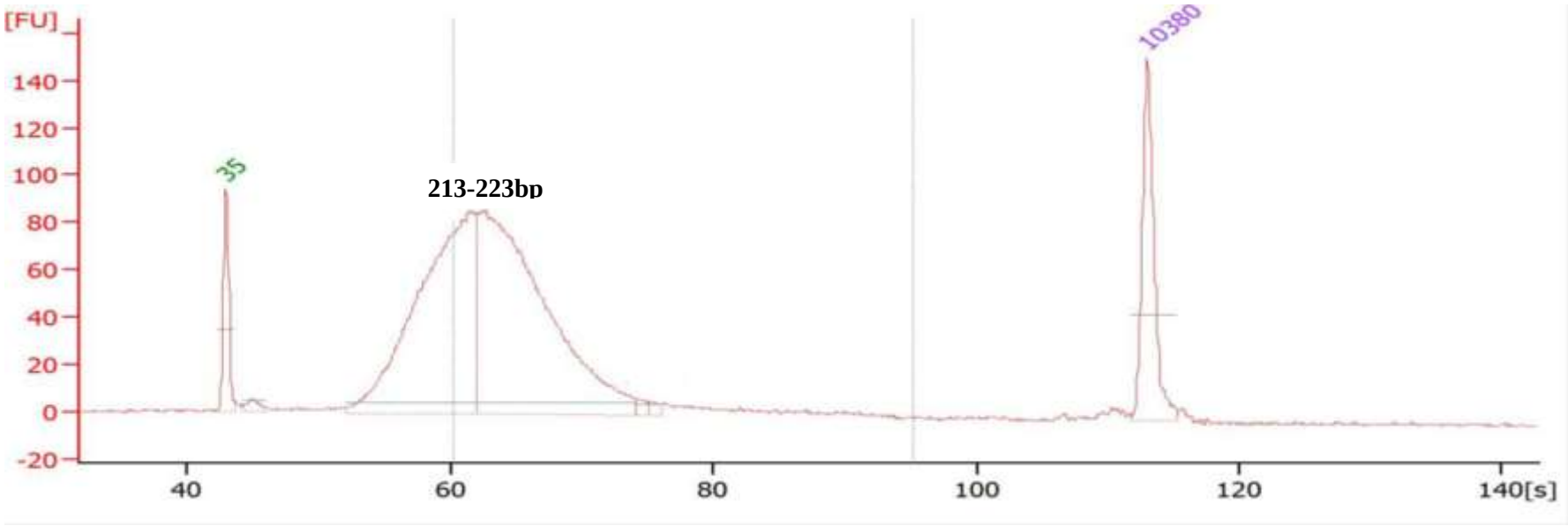
Figure 3.6. three electropherograms of the same *FY* DNA library.

A) 5 minute-fragmented library with the peak at around size 200-300bp.

B) Same library was ligated with clear shift to the right.

C) Size election was done by Pippin Prep™ instrument targeting the size around 300bp.

D) The red circle represents the concentration calculated to determine the Dilution factor for next step (template preparation) here the total is 1662.3 pmol/l. The green line (35bp) is the lower marker and the purple line (10380bp) is the upper marker. Results shown were obtained using the Bioanalyzer® instrument.



Peak		Size [bp]	Conc. [pg/µl]	Molarity [pmol/l]	Observations
1	◀	35	125.00	5,411.3	Lower Marker
2		48	15.78	500.9	
3		213	628.46	4,471.8	
4		223	732.48	4,970.2	
5		360	7.25	30.5	
6		373	4.49	18.2	
7	▶	10,380	75.00	10.9	Upper Marker

Figure 3.7 an electropherograms of size selected *FY* DNA library (by SPRIselect® reagent magnetic bead).

It can be seen that peak is around 200bp (212-223bp). The red circle represents the concentration calculated to determine the Dilution factor for next step (template preparation) here the total is 9442 pmol/l. The green line (35bp) is the lower marker and the purple line (10380bp) is the upper marker. Results shown were obtained using the Bioanalyzer® instrument.

3.3.3 Next Generation Sequencing data quality control

3.3.3.1 Sequencing Data summary report

Following sequencing, the sequencing data were transferred to the Ion Torrent Server software and was processed, for example, by trimming adapters' sequence. A sequencing summary of the data generated is by default produced by Torrent Suite™ Software Version 4.4. The average of the total reads generated from the 53 *FY* samples was ~ 3.130 million, with a mean coverage depth of about 5600x. Table 3.2 summarises the sequencing report of the total reads of the *FY* samples from the 3 separate runs). Figure 3.8 illustrates a part of a sequencing report of single *FY* sequencing run on the Ion PGM™, which is representative of other runs. Here, the loading percentage of the 316™ chip wells was 75%, which indicates the potential addressable wells containing ISPs. The total number of usable reads for downstream analysis was ~3.5 million, which made up 74% of the total reads containing library ISPs. These reads were processed in two steps using software to ensure their quality and accuracy: well classification and read filtering (and trimming). The former step mainly distinguishes the empty wells from the loaded wells, with the ISP-addressable wells providing the bead loading density (ISP loading percentage). In addition, this step distinguishes between control fragments and the library ISPs. The second step is applied to trim the reads from polyclonal ISPs (22%), while maintain those clonal ISPs (78%). Clonal ISPs represent a population of single unique template fragments that have been amplified by Ion OneTouch™, whereas the polyclonal reads represent fragments containing more than just one amplified sequence. Moreover, reads of low quality, short reads (less than 4 base pairs) and primer dimers (less than 8 bp) are filtered out. The percentage of the final library was 94%, with a mean read length of 143 bp.

Following this, different aspects of the sequencing data were analysed by various plugins of the Torrent Suite™, generating Variant Call format (VCF) files and Binary Alignment/Map (BAM) files. The quality of the sequencing reads was checked by FastQC. Variants were analysed by the variantCaller plugin, which was linked to IGV for visualisation.

Table 3.2. A summary of the Ion PGM™ sequence run output of the 53 *FY* samples, collectively processed in 3 runs.

	No. of samples genotyped	**ISP Loading %**	**Total usable reads**	**Usable reads%**	**Mean read length**
1st run report	12	71	3,007,198	69	167 bp
2nd run report	17	67	2,884,413	69	164 bp
3rd run report	24	75	3,501,099	74	143bp

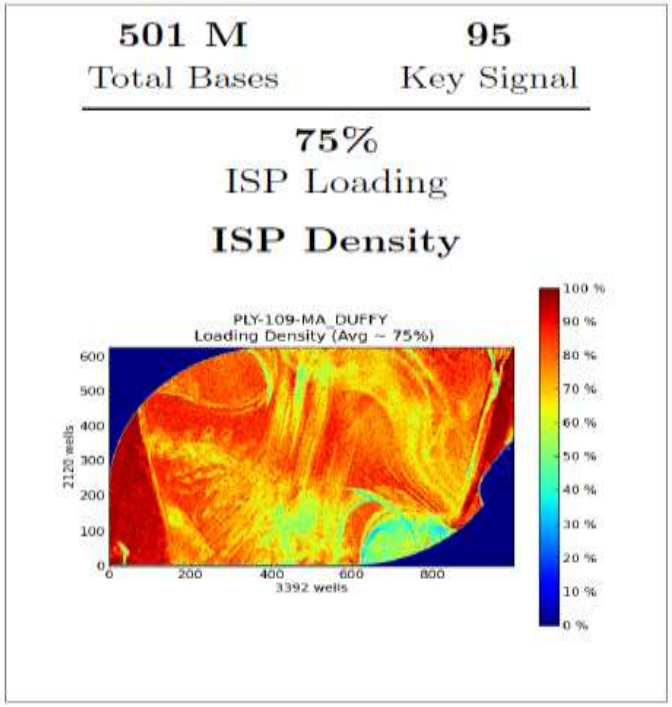
Figure 3.8 A summary report for a single sequencing run of *FY* library. *(NOTE: the figure is on the next page).*

A) 501 million bp were aligned to the reference. The percentage (75%) represents chip wells containing the ISPs (both templated and non-templated). The colour represents the loading percentage of ISP across the physical 316^{TM} chip plate surface. 99% of filtered bases of all reads were aligned to the reference.

B) The total number of reads is 3,501,099, which are provided after trimming and filtration from empty wells, non-templated and polyclonal reads. The percentage of usable sequence (74%) is calculated by dividing the percentage of filtered library reads (3,501,099)/ number of the reads containing ISPs (4,735,364) multiplied by 100. The live/enrichment percentage is 100%, which indicates that ISPs contain a strong sequence signal from test fragment and library.

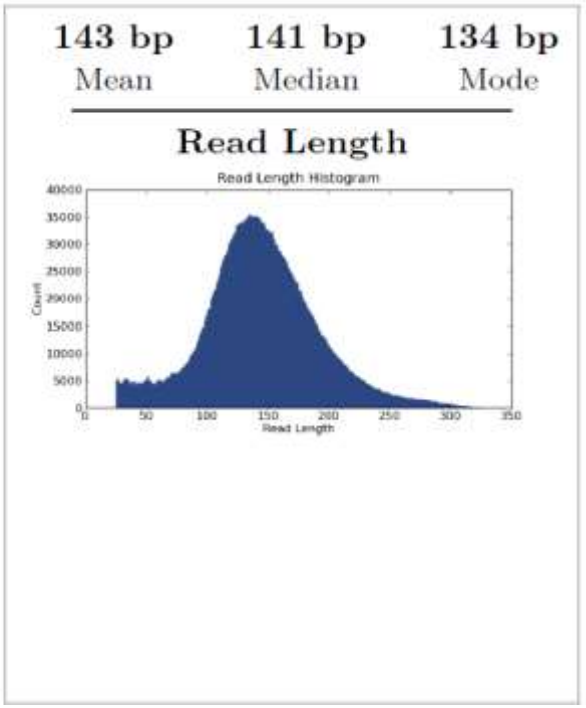
C) A histogram shows a mean reading length of 143bp. The read count is displayed in the y-axis, while the read length, in bp, is shown on the x-axis.

A)



B)

C)



Results of Well classification

Addressable Wells	**6,337,389**	
With ISPs	4,779,287	75.4%
Live	4,766,148	99.7%
Test Fragment	30,784	00.6%
Library	4,735,364	99.4%

Results of filtering and trimming

Library ISPs	**4,735,364**	
Filtered: Polyclonal	1,041,624	22.0%
Filtered: Low Quality	181,853	03.8%
Filtered: Adapter Dimer	10,788	00.2%
Final Library ISPs	**3,501,099**	**73.9%**

3.3.3.2 NGS data quality control

The quality of the NGS data is crucial for further analysis, which includes genotyping. The mean coverage depth (5600X) (obtained by using coverage analysis: an Ion Torrent Suite™ plugin), which is the number of times the reads and bases were sequenced and aligned to the reference sequence (which increases the confidence for variant calling) and various quality aspects were analysed. FastQC, a plugin on the Torrent Suite, was used to check the quality of the generated sequence data as 'per base sequence quality' and 'per sequence quality' scores, with reference to the Phred score, which gives an initial impression of the quality and status of the data. The FastQC quality evaluation of the data is explained in the following two sections.

3.3.3.2.1 Per base sequence quality

Figure 3.9 illustrates an overview of the quality of sequenced base pairs at their position on the reads, according to Phred score (Table 3.3). In this run, the mean quality score of sequenced bases was around 30-31 (Phred score represented in the y axis). As a result, according to the Phred score, the base call accuracy is 99.9% – a probability of 1 in 1000 of the base being incorrectly called. To explain the graph briefly, the background of the graph splits the y axis into three coloured parts, describing the quality scores: the best quality call (top/green), reasonable (middle/orange) and poor quality (bottom/red). A BoxWhisker-type plot is drawn for each base pair position, which itself consists of number of parts. Yellow boxes represent the inter-quartile range (25-75%), within which the median value is illustrated as a central red line. The mean quality value is represented by a blue line that, here, mainly indicates a value of (30-31) but reduces when approaching the longer read length. The reason for this is that the quality of calls of this platform, and most platforms, drops as the run approaches the end of the reads, which is due to the degradation of sequencing chemistry (Andrews, 2016). The 10% and 90% points are represented by the upper and lower whiskers. The quality of all three

runs was comparable (above 99% base call accuracy, according to the Phred score; Table 3.4).

3.3.3.2.2 Per sequence quality scores

An output report from the FastQC plugin demonstrates the quality score, according to the Phred score, of the sequences and assesses whether a subset of these were generated with low quality values. The sequences reads shown in Figure 3.10 mainly received a 30 Phred score, which matches the quality indication previously discussed. Accordingly, these parameters and the coverage depth provide confidence in the high throughput data for further analysis. The quality of all three runs was comparable (above 99% base call accuracy, according to the Phred score; Table 3.4).

Table 3.3 Phred quality score and the base call accuracy. (Ion Torrent community)
The various Phred scores with their quality representation for base call accuracy are listed.

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

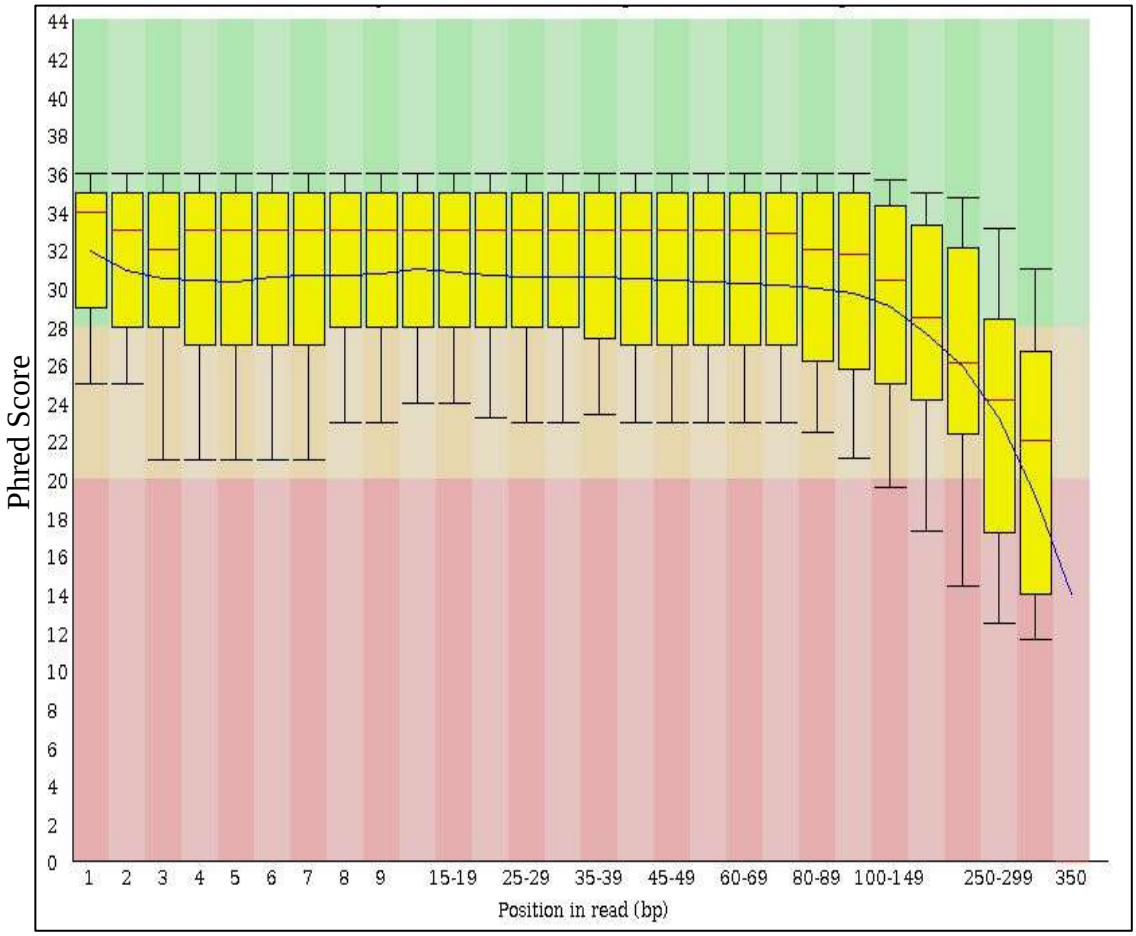


Figure 3.9 Mean Phred quality scores across all bases of *FY* samples in a single run.

The Phred score is displayed on the y-axis, whereas the x-axis represents the position of bases in the reads. Three colours of background represent levels of quality, according to the Phred score: very good (green), reasonable (orange) and poor quality base calling (red). A BoxWhisker-type plot is drawn for each position with the 25-75% interquartile range represented by yellow boxes. Upper and lower whiskers represent the 10% and 90% points. The blue line represents the mean value of the base call quality (30-31), which indicates a 99.9% base call accuracy. The red line denotes the median value of the quality. The other *FY* runs showed a comparable output.

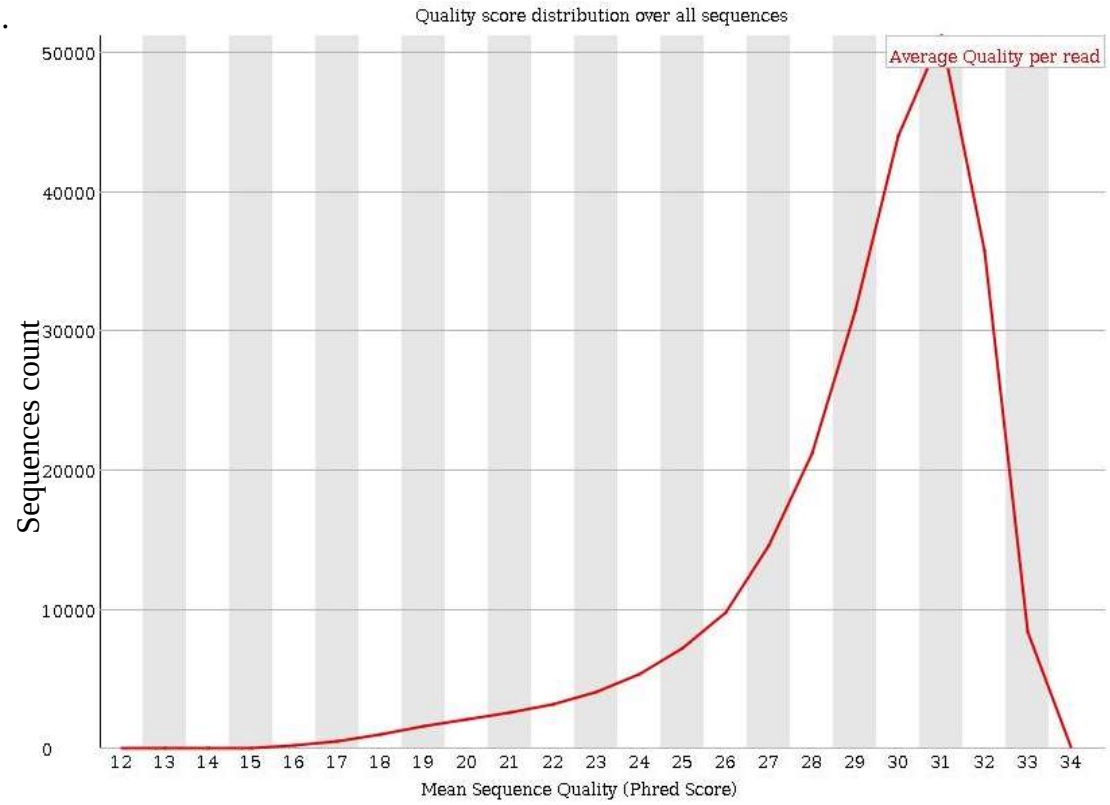


Figure 3.10 The mean quality score of the *FY* sequences generated from a single run.

The mean quality score (Phred score, x-axis) of the sequences across the number of reads (y-axis) is shown. The majority of sequences have a mean quality of more than 30, which indicates high quality of generated reads with an accuracy of 99.9% and a probability of 1 in 1000 that the base was incorrectly called. Other runs had comparable results.

Table 3.4 Summary of the sequencing quality of the 3 *FY* NGS runs.

All runs showed base call accuracy above 99%, especially the third run, which has a quality of 99.9%.

Run	Per base sequence quality	Per sequencing quality score
1st	27-28	26-27
2nd	28-30	28
3rd	30-31	30-31
Base call accuracy	above 99%	above 99%

3.3.3.3 NGS sequence visualisation

Different aspects of the 53 samples of *FY* complete gene sequencing were analysed. The sequencing data was visualised by the IGV software (2.3) using the BAM files. Using this software, the sequencing reads were aligned against the reference *FY* gene sequence of human genome (hg19), allowing assessment of various crucial aspects of the data. The nucleotide and amino acid locations were based on the major *FY* transcript that consists of 2 exons with associated mRNA (NM_002036). The integrity of the sequence, in terms of the full coverage of the *FY* gene as well as the flanking regions, coverage depth and genomic variations are visible. Examples of genomic information obtained from IGV are: gene mutations, such as SNPs, insertions and deletions; zygosity; and the chromosomal location of the nucleotides. All these data was assessed by IGV and at least one other software package or database in order to ensure validity and accuracy of the variants detected (Robinson et al., 2011). The reference sequence of the *FY* gene is also shown as nucleotides and the amino acids sequence it encodes. Figure 3.11 illustrates an analysis, using IGV, to visualise and assess a *FY* sample with serologically defined phenotype Fy (a+b+). The distribution of the sequence across the whole gene plus up- and downstream flanking regions is shown, reflecting the integrity of the library preparation step and the specificity of the designed primers. In addition, in this figure the *FY*A*/*FY*B* critical missense mutation (SNP) (125G>A, in exon 2) at chromosomal location ch1: 159175354 was shown to be heterozygous (50% of each G and A were represented), with a coverage depth at this base of 6116X. As a result, the amino acid substitution Gly42Asp occurred. Two intronic SNPs were found in intron 1: G>C 159174824 and C>T 159174920; homozygous and heterozygous, respectively. Accordingly, genotyping of this SNP in exon 2 matched the serological phenotype provided with the sample (*FY*A/FY*B* and Fy a+b+).

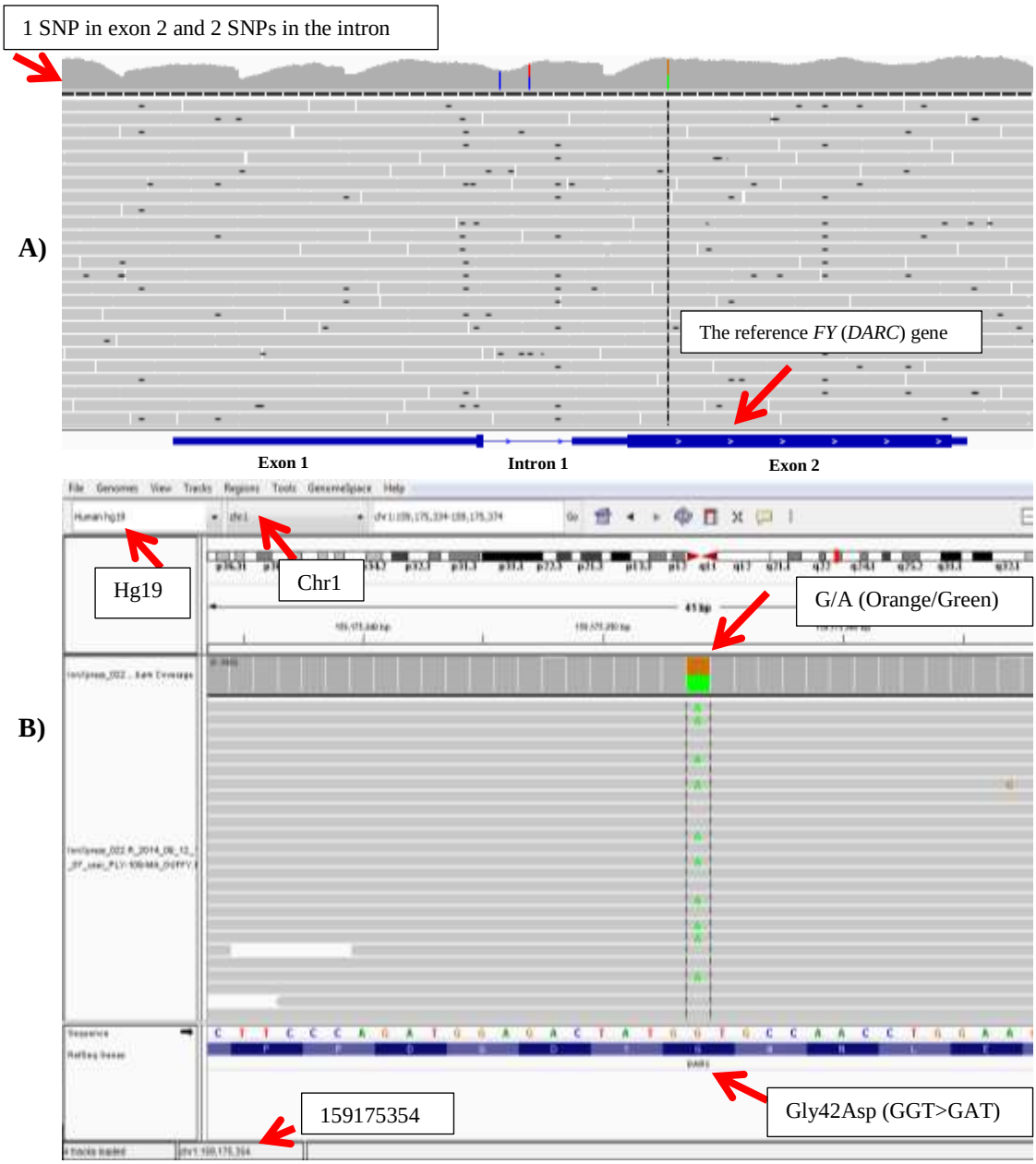


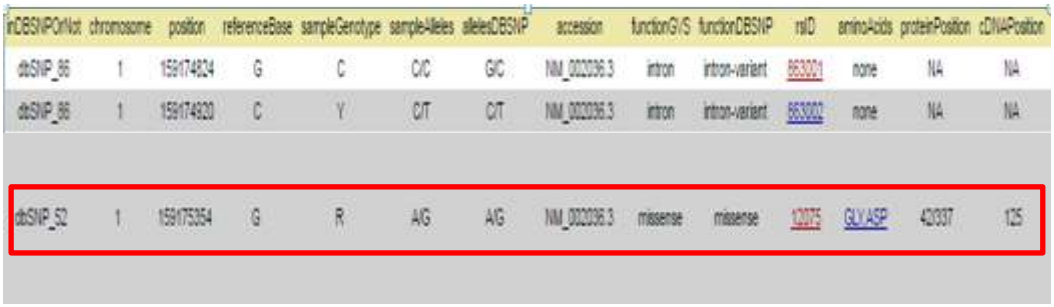
Figure 3.11 An IGV image illustrating the visualisation analysis of the sequencing data of a Fy (a+b+) sample with *FY*A*/*FY*B* genotype.

A) The reference *FY* (DARC) gene (blue) is fully covered and aligned by the sequencing reads. The coverage depth is shown above. The coloured bars above denote SNPs: 1 SNP in exon 2 and 2 in intron 1(C>T 159174920 and G>C 159174824). The first intronic SNP is homozygous (solid bar) while the latter is heterozygous (split bar).

B) The missense mutation (SNP 125G>A, in exon 2), at chromosomal position 159175354 encodes an amino acid substitution of Gly42Asp (GGT>GAT). This SNP is heterozygous, displayed by the spilt box (orange/green) for 50% G and 50% A, respectively. This shows a match between the genotype and serology (*FY*A*/*FY*B* and Fy a+b+).

3.3.3.4. Variant analysis and Genotyping

The sequencing data generated from the Ion PGM™ and projected from Ion Suite were analysed for various mutations, such as SNPs (in exons, intron and flanking regions), insertions and deletions, in addition to zygosity. Firstly, the Variant Caller software, which is a plugin of the Ion Suite™ software, was used, which analyses the aligned reads generated in BAM format generated from the Ion Suite™. This programme has an advantage over other annotation software packages as it can be set to automatically analyse the sequencing data in terms of mutations annotations and zygosity against the reference gene. In addition, this software is linked to the IGV software, enabling direct visualisation of the mutations and the ability to conduct further analysis (see section 3.3.3.3). Moreover, the Data generated from the Variant Caller are in VCF format files which were used with different party software for confirmation. The VCF files were utilised and uploaded in an annotation online tool SeattleSeq Annotation 137, which provides a wide range of information, such as the chromosomal location mutations, type of mutation, the transcript number, amino acid change and nucleotides the cDNA based on database in the NCBI. Figure 3.12 illustrates an example of the SeattleSeq Annotation 137 output.



inDBSNP/Not	chromosome	position	referenceBase	sampleGenotype	sampleAlleles	alleleDBSNP	accession	functionG/S	functionDBSNP	rsID	aminoAcids	proteinPosition	cDNAPosition
dbSNP_86	1	159174824	G	C	C/C	G/C	NM_002036.3	intron	intron-variant	663021	none	NA	NA
dbSNP_86	1	159174820	C	Y	C/T	C/T	NM_002036.3	intron	intron-variant	663022	none	NA	NA
dbSNP_52	1	159175354	G	R	A/G	A/G	NM_002036.3	missense	missense	12075	GLY:ASP	42/337	125

Figure 3.12 An example of a variant annotation report provided from the SeattleSeq Annotation 137 online tool of a Fy (a+b+) sample with *FY*A*/*FY*B* genotype.

As an example, the highlighted mutation (*FY*A*/*FY*B* critical missense mutation) gives information on SNPs, chromosomal location (ch1:159175354), the reference base (G), heterozygosity (A/G located at cDNA 125), reference transcript number (NM_002036.3) and a missense mutation encoding Gly42Asp. The SNP is known as has been reported in the dbSNP and has an rsID number. Two intronic SNPs are also shown.

3.3.4 *FY* genotyping results from Next Generation Sequencing

NGS provided an extensive genotyping of the entire *FY* gene and flanking regions to reveal all existing mutations (known and novel). Sequencing and analysis of 53 individual samples revealed various mutations, including SNPs. The genotype was then correlated with the phenotype.

3.3.4.1 Mutations in exons and the promoter region

43 out of 53 samples were of a known FY serological phenotype. The NGS genotyping data matched and confirmed these phenotypes with *FY* alleles *FY*A*, *FY*B* and *FY*02*(*Null*) main polymorphisms. The SNPs found in exons and the promoter region are listed in Table 3.5. The crucial SNP *FY*A*/*FY*B* (125G>A) in exon 2, encoding for Gly42Asp, was in concordance with FY phenotype and zygosity. The silencing SNP (-67 T>C) in the promoter region was found to be homozygous in 7 out of 53 Fy (a-b-) samples carrying a *FY*B* background with homozygous 125G>A encoding for the *FY*02N.01* allele. 16 out of 53 samples were found to carry the SNP 298G>A (15 were heterozygous and 1 homozygous) encoding for the amino acid change Ala100Thr. This SNP appears to be associated with the *FY*B* allele as it was found to be homozygous in a *FY*B*/*FY*B* sample, 15 heterozygous situations across (9 *FY*B*/*FY*B,* 5 *FY*A*/*FY*B* and 1 *FY*02M.01)* and was homozygous G in *FY*A/ FY*A*. All phenotyped samples carrying this SNP show positive expression of Fy^b, suggesting that this mutation has no effect on Fy^b antigenicity. In addition, 1 sample of *FY*B* background carried 2 SNPs: 298G>A (Ala100Thr), along with a 265C>T point mutation that encodes the amino acid change Arg89Cys. The sample phenotype was Fy (a-b+), suggesting normal expression of Fy^b, although the 2 SNPs combined relate to the phenotype Fy^x and *FY*02M.01*allele. Only 1 out of 53 samples, of phenotype Fy (a+b+), was shown to carry a SNP (714G>A) in exon 2 encoding a synonymous substitution Gly238Gly. Accordingly, the NGS genotype was *FY*A/ FY*B*. Table 3.7 provides more detailed information regarding the chromosomal location of all the discussed SNPs. The confirmed determination of those SNPs, for example 265C>T and 298G>A whether carried on the same allele (*cis*) or not (*trans*) might be challenging by the NGS platform. This is because of the chemistry of the sequencing that needs small fragments 200-400bp (see section 6.4).

These SNPs observed from NGS genotyping required no further confirmation or validation because as well as the fact that the coverage depth for all the *FY* samples was significantly high (5600X), all these SNPs are reported in the NCBI database (dbSNP) and have been previously validated.

3.3.4.2 Mutations in the intron and downstream region.

As the NGS protocol here was optimised to cover the entire *FY* gene and flanking region, this allowed extensive exploration of all existing gene mutations, including those in introns and outer regions. Four SNPs and 1 deletion were found in intron 1 (Table 3.6). From Table 3.7, it can be seen that there is no complete correlation between *FY*A*, *FY*B* and *FY*02 Null* alleles and intronic SNPs; although, SNP 159174885 T>C was shown to be heterozygous and homozygous in *FY*B* alleles and absent in *FY*A* alleles. One SNP (G>C 159174824) was found to be homozygous in all 53 samples (all *FY* alleles), while the deletion (T>Del 159175005) was homozygous in all *FY*B* and *FY*Null* alleles. However, this deletion was also found in 14 *FY*A* haplotypes (8 heterozygous and 1 homozygous in *FY*A*, *FY*A*, and 4 homozygous in *FY*A*, *FY*B*). Several SNPs were found in up- and downstream regions, at various distances, and, although these were found on the NCBI database, their effect on the *FY* gene remains unknown. Despite the fact there is no complete correlation of intronic SNPs with *FY*A*, *FY*B* and *Null* alleles, the analysis of the distribution of these SNPs illustrated unique patterns that may define new *FY* alleles (Table 3.7), as various sets of these intronic SNPs present differently in samples of the same phenotype.

			NGS defined sequence				
NGS Genotype (Alleles*)	**Number of samples**	**Variation**					**Serology (n)**
		(Exon 2) ***FY*A / FY*B*** **125G>A Gly42Asp**	**(Exon2) 265C>T Arg89Cys**	**(Exon2) 298G>A Ala100Thr**	**(Exon2) 714G>A Gly238Gly**	**(promoter) (- 67 T>C)**	
*FY*A/FY*A*	16	G/G (16)	C/C (16)	G/G (16)	G/G (16)	T/T (16)	Fy (a+b-) (15)
*FY*B/FY*B*	16	A/A (16)	C/C (16)	G/G (6) G/A (9) A/A (1)	G/G (16)	T/T (16)	Fy (a-b+) (13)
*FY*A/ FY*B*	13	G/A (13)	C/C (13)	G/G (8) G/A (5)	G/G (12) G/A (1)	T/T (13)	Fy (a+b+) (7)
*FY*B/FY*02M.01*	1	A/A (1)	C/T (1)	G/A (1)	G/G (1)	T/T (1)	Fy (a-b+) (1)
*FY*02N.01/ FY*02N.01*	7	A/A (7)	C/C (7)	G/G (7)	G/G (7)	C/C (7)	Fy (a-b-) (7)

Table 3.5 NGS genotyping of 53 samples of differing Fy phenotypes

53 different DNA samples (43 of which were phenotyped) were sequenced by *FY*-specific LR-PCR. All phenotyped samples showed a complete concordance with NGS genotyping data, particularly 125G>A and T-67C mutations. 16 out of 53 samples were shown to carry the 298G>A mutation (15 heterozygous and 1 homozygous) with a positive phenotype of Fy^b – suggesting normal Fy^b antigenicity. 1 Fy (a-b+) sample was shown to be heterozygous for both 298G>A and 265C>T (Arg89Cys), associated with the Fy^x phenotype. SNP 714G>A in exon 2 was found in one Fy (a+b+) sample.* Allele names are the same as in the ISBT (Reid et al., 2012).

Table 3.6 The list of SNPs in the intron of the *FY* gene and flanking regions.

The table shows information on the chromosomal position, the base change, and the position in and from the gene. *Distance from the last nucleotide of exon 1 (+) whereas (-) from the first nucleotide of exon 2 (only those in introns is provided). These mutations are reported in the NCBI (dbSNP) database and have a rsID (rs number). Grey highlighted polymorphisms have been described and studied for the *FY* alleles before (Schmid et al., 2012).

rsID	Chromosomal position	Position in the Gene	Nucleotide change	IVS1*
rs3027009	159173887	-863 upstream	A>G	-
rs41264467	159174077	-673 upstream	G>A	-
rs3027012	159174123	-627 upstream	C>T	-
rs3027013	159174209	-541 upstream	C>T	-
rs2814778	**159174683**	**Promoter**	**- 67 T>C**	**-**
rs863001	**159174824**	**Intron 1**	**G>C**	**+54**
rs7550207	**159174885**	**Intron 1**	**T>C**	**+115**
rs863002	**159174920**	**Intron 1**	**C>T**	**+150**
rs17838198	**159175005**	**Intron 1**	**T>DEL**	**-246**
rs3027016	**159175193**	**Intron 1**	**A>G**	**-58**
rs12042349	159176490	250 Downstream	C>T	-
rs863003	159176508	268 Downstream	A>G	-

Table 3.7 NGS genotyping of 53 *FY* samples, 43 of which are of known phenotype *(NOTE: the table is on the next page).*

The table graphically illustrates the various *FY* alleles found in this study. The SNPs in *FY*A* (the reference sequence, according to the 125G>A change found in the NCBI database) are shown in red, blue in *FY*B* and green in *FY*0(Null)*. Homozygous mutations are represented by solid colour, whilst crossed colour represents heterozygous mutations. The chromosomal position (chromosome 1/hg19), SNP location in the *FY* gene and amino acid changes are shown at the top of the table. The *FY*A/FY*B* SNP is highlighted in orange, SNPs in exons and the promoter region are yellow, high frequency SNPs (such as the homozygous 159174824) are grey. Sample number and phenotype are shown on the left side of the table. Analysis of the exonic and intronic SNPs revealed different patterns of *FY* alleles: 4 *FY*A* alleles, more than 4 *FY*B* and 2 *FY*02N.01.* Samples with no phenotype were denoted as ND (not defined). 2 samples were of Fy (a+) but not (b) status, so the phenotype for these samples is denoted as Fy(a+b ?*). SNPs in up- and downstream flanking regions are also listed. The number 008 was used in accordance with the number of the FY blood group system (ISBT number). The rsID for polymorphisms are shown on the top row. Larger table in appendix B.

		rsID	3027009	41264467	3027012	3027013	2814778	863001	7550207	863002	17838198	3027016	12075	34599082	13962	36007769	12042349	863003
			A>G	G>A	C>T	C>T	(- 67 T>C	G>C	T>C	C>T	T>DEL	A>G	125G>A	265 C>T	298 G>A	714G>A	C>T	A>G
			159173887	159174077	159174123	159174209	159174683	159174824	159174885	159174920	159175005	159175193	159175354	159175494	159175527	159175943	159176490	159176508
			Upstream	Upstream	Upstream	Upstream	Promoter	intron 1	intron 1	intron 1	intron 1	intron 1	Exon 2	Exon2	Exon2	Exon2	Downstream	Downstream
											HIGH FREQUENCY		*FY*A/FY*B* GLY42ASP	Arg89Cys	Ala100Thr	Gly238Gly		
Genotype																Synonymous variant		
*16 FY*A/FY*A*	SAMPLE NUMBER	Phenotype																
*FY*A/FY*A*	FY008.01	ND	A/A	G/G	C/C	C/C	T/T	C/C	T/T	C/T	T/DEL	A/A	G/G	C/C	G/G	G/G	C/C	A/A
*FY*A/FY*A*	FY008.02	Fy (a+b-)	A/A	G/G	C/C	C/C	T/T	C/C	T/T	C/C	T/T	A/A	G/G	C/C	G/G	G/G	C/C	A/A
*FY*A/FY*A*	FY008.03	Fy (a+b-)	A/A	G/G	C/C	C/C	T/T	C/C	T/T	C/C	T/DEL	A/A	G/G	C/C	G/G	G/G	C/T	A/A
*FY*A/FY*A*	FY008.04	Fy (a+b-)	A/A	G/G	C/C	C/C	T/T	C/C	T/T	C/C	T/DEL	A/A	G/G	C/C	G/G	G/G	C/T	A/A
*FY*A/FY*A*	FY008.05	Fy (a+b-)	A/A	G/G	C/C	C/C	T/T	C/C	T/T	C/C	T/DEL	A/A	G/G	C/C	G/G	G/G	C/T	A/A
*FY*A/FY*A*	FY008.06	Fy (a+b-)	A/A	G/G	C/C	C/C	T/T	C/C	T/T	C/C	T/DEL	A/A	G/G	C/C	G/G	G/G	C/T	A/A
*FY*A/FY*A*	FY008.07	Fy (a+b-)	A/A	G/G	C/C	C/C	T/T	C/C	T/T	C/C	T/DEL	A/A	G/G	C/C	G/G	G/G	C/T	A/A
*FY*A/FY*A*	FY008.08	Fy (a+b-)	A/A	G/G	C/C	C/C	T/T	C/C	T/T	C/C	T/T	A/A	G/G	C/C	G/G	G/G	C/C	A/A
*FY*A/FY*A*	FY008.09	Fy (a+b-)	A/A	G/G	C/C	C/C	T/T	C/C	T/T	C/C	T/DEL	A/A	G/G	C/C	G/G	G/G	C/T	A/A
*FY*A/FY*A*	FY008.10	Fy (a+b-)	A/A	G/G	C/C	C/C	T/T	C/C	T/T	C/T	T/DEL	A/A	G/G	C/C	G/G	G/G	C/C	A/A
*FY*A/FY*A*	FY008.11	Fy (a+b-)	A/A	G/G	C/C	C/C	T/T	C/C	T/T	C/T	DEL/DEL	A/A	G/G	C/C	G/G	G/G	C/T	A/A
*FY*A/FY*A*	FY008.12	Fy (a+b-)	A/A	G/G	C/C	C/C	T/T	C/C	T/T	C/C	T/T	A/A	G/G	C/C	G/G	G/G	C/C	A/A
*FY*A/FY*A*	FY008.13	Fy (a+b-)	A/A	G/G	C/C	C/C	T/T	C/C	T/T	C/C	T/T	A/A	G/G	C/C	G/G	G/G	C/C	A/A
*FY*A/FY*A*	FY008.14	Fy (a+b-)	A/A	G/G	C/C	C/C	T/T	C/C	T/T	C/C	T/T	A/A	G/G	C/C	G/G	G/G	C/C	A/A
*FY*A/FY*A*	FY008.15	Fy (a+b-)	A/A	G/G	C/C	C/C	T/T	C/C	T/T	C/C	T/T	A/A	G/G	C/C	G/G	G/G	C/C	A/A
*FY*A/FY*A*	FY008.16	Fy (a+b-)	A/A	G/G	C/C	C/C	T/T	C/C	T/T	C/C	T/T	A/A	G/G	C/C	G/G	G/G	C/C	A/A
*16 FY*B/FY*B*																		
*FY*B/FY*B*	FY008.17	Fy (a-b+)	A/G	G/G	C/T	C/T	T/T	C/C	C/T	C/T	DEL/DEL	A/A	A/A	C/C	G/G	G/G	C/C	A/G
*FY*B/FY*B*	FY008.18	Fy (a-b+)	A/A	G/G	C/C	C/C	T/T	C/C	C/T	C/T	DEL/DEL	A/G	A/A	C/C	G/A	G/G	C/C	A/G
*FY*B/FY*B*	FY008.19	ND	A/A	G/G	C/C	C/C	T/T	C/C	T/T	T/T	DEL/DEL	A/A	A/A	C/C	A/A	G/G	C/C	G/G
*FY*B/FY*B*	FY008.20	ND	A/A	G/G	C/C	C/C	T/T	C/C	T/T	T/T	DEL/DEL	A/A	A/A	C/C	G/A	G/G	C/C	G/G
*FY*B/FY*B*	FY008.21	ND	A/A	G/G	T/T	C/C	T/T	C/C	T/T	T/T	DEL/DEL	A/A	A/A	C/C	G/G	G/G	C/C	A/A
*FY*B/FY*B*	FY008.22	Fy (a-b+)	A/A	G/G	C/T	C/C	T/T	C/C	T/T	T/T	DEL/DEL	A/A	A/A	C/C	G/A	G/G	C/C	A/G
*FY*B/FY*B*	FY008.23	Fy (a-b+)	G/G	G/G	T/T	T/T	T/T	C/C	C/C	C/C	DEL/DEL	A/A	A/A	C/C	G/G	G/G	C/C	A/A
*FY*B/FY*B*	FY008.24	Fy (a-b+)	A/G	G/G	C/T	C/T	T/T	C/C	C/C	C/C	DEL/DEL	A/G	A/A	C/C	G/G	G/G	C/C	A/A
*FY*B/FY*B*	FY008.25	Fy (a-b+)	A/A	G/G	C/C	C/C	T/T	C/C	C/T	C/T	DEL/DEL	A/G	A/A	C/C	G/A	G/G	C/C	A/G
*FY*B/FY*B*	FY008.26	Fy (a-b+)	A/A	G/G	C/C	C/C	T/T	C/C	T/T	T/T	DEL/DEL	A/A	A/A	C/C	G/A	G/G	C/C	G/G
*FY*B/FY*B*	FY008.27	Fy (a-b+)	A/G	G/G	C/T	C/T	T/T	C/C	C/T	C/T	DEL/DEL	A/A	A/A	C/C	G/A	G/G	C/C	A/G
*FY*B/FY*B*	FY008.28	Fy (a-b+)	A/G	G/G	T/T	C/T	T/T	C/C	C/T	C/T	DEL/DEL	A/A	A/A	C/C	G/G	G/G	C/C	A/A
*FY*B/FY*B*	FY008.29	Fy (a-b+)	A/A	G/G	C/C	C/C	T/T	C/C	T/T	T/T	DEL/DEL	A/A	A/A	C/C	G/A	G/G	C/C	G/G
*FY*B/FY*B*	FY008.30	Fy (a-b+)	A/G	G/A	C/T	C/T	T/T	C/C	C/C	C/C	DEL/DEL	A/G	A/A	C/C	G/G	G/G	C/C	A/A
*FY*B/FY*B*	FY008.31	Fy (a-b+)	A/G	G/G	C/T	C/T	T/T	C/C	C/T	C/T	DEL/DEL	A/A	A/A	C/C	G/A	G/G	C/C	A/G
*FY*B/FY*B*	FY008.32	Fy (a-b+)	A/A	G/G	C/T	C/C	T/T	C/C	T/T	T/T	DEL/DEL	A/A	A/A	C/C	G/A	G/G	C/C	A/G
*1 FY*B/FY*02M.01*																		
FY B/FY*02M.01*	FY008.33	Fy (a-b+)	A/A	G/G	C/T	C/C	T/T	C/C	T/T	T/T	DEL/DEL	A/A	A/A	C/T	G/A	G/G	C/C	A/G
7 FY 02N.01/FY* 02N.01*																		
FY 02N.01/FY* 02N.01*	FY008.34	Fy (a-b-)	A/A	G/G	C/C	C/C	C/C	C/C	T/T	C/C	DEL/DEL	A/A	A/A	C/C	G/G	G/G	C/C	A/A
FY 02N.01/FY* 02N.01*	FY008.35	Fy (a-b-)	A/A	G/G	C/C	C/C	C/C	C/C	T/T	C/C	DEL/DEL	A/A	A/A	C/C	G/G	G/G	C/C	A/A
FY 02N.01/FY* 02N.01*	FY008.36	Fy (a-b-)	A/A	G/G	C/C	C/C	C/C	C/C	T/T	C/C	DEL/DEL	A/A	A/A	C/C	G/G	G/G	C/C	A/A
FY 02N.01/FY* 02N.01*	FY008.37	Fy (a-b-)	A/A	G/G	C/C	C/C	C/C	C/C	C/C	C/C	DEL/DEL	A/A	A/A	C/C	G/G	G/G	C/C	A/A
FY 02N.01/FY* 02N.01*	FY008.38	Fy (a-b-)	A/A	G/G	C/C	C/C	C/C	C/C	T/T	C/C	DEL/DEL	A/A	A/A	C/C	G/G	G/G	C/C	A/A
FY 02N.01/FY* 02N.01*	FY008.39	Fy (a-b-)	A/A	G/G	C/C	C/C	C/C	C/C	T/T	C/C	DEL/DEL	A/A	A/A	C/C	G/G	G/G	C/C	A/A
FY 02N.01/FY* 02N.01*	FY008.40	Fy (a-b-)	A/A	G/G	C/C	C/C	C/C	C/C	T/T	C/C	DEL/DEL	A/A	A/A	C/C	G/G	G/G	C/C	A/A
*13 FY*A/FY*B*																		
*FY*A/FY*B*	FY008.41	Fy (a+b?*)	A/A	G/G	C/C	C/C	T/T	C/C	T/T	C/T	T/DEL	A/A	G/A	C/C	G/G	G/G	C/C	A/G
*FY*A/FY*B*	FY008.42	Fy (a+b+)	A/A	G/G	C/C	C/C	T/T	C/C	T/T	C/T	T/DEL	A/A	G/A	C/C	G/A	G/G	C/C	A/G
*FY*A/FY*B*	FY008.43	ND	A/G	G/G	C/T	C/T	T/T	C/C	C/T	C/C	T/DEL	A/A	G/A	C/C	G/G	G/G	C/C	A/A
*FY*A/FY*B*	FY008.44	ND	A/A	G/G	C/C	C/C	T/T	C/C	T/T	T/T	DEL/DEL	A/A	G/A	C/C	G/G	G/G	C/C	A/G
*FY*A/FY*B*	FY008.45	ND	A/A	G/G	C/C	C/C	T/T	C/C	C/T	C/C	T/DEL	A/G	G/A	C/C	G/G	G/G	C/C	A/A
*FY*A/FY*B*	FY008.46	Fy (a+b?*)	A/A	G/G	C/T	C/C	T/T	C/C	T/T	C/T	DEL/DEL	A/A	G/A	C/C	G/G	G/G	C/T	A/A
*FY*A/FY*B*	FY008.47	ND	A/A	G/G	C/C	C/C	T/T	C/C	T/T	C/T	T/DEL	A/A	G/A	C/C	G/A	G/G	C/C	A/G
*FY*A/FY*B*	FY008.48	Fy (a+b+)	A/A	G/G	C/C	C/C	T/T	C/C	T/T	C/T	T/DEL	A/A	G/A	C/C	G/A	G/G	C/C	A/G
*FY*A/FY*B*	FY008.49	Fy (a+b+)	A/A	G/G	C/C	C/C	T/T	C/C	C/T	C/C	T/DEL	A/G	G/A	C/C	G/G	G/G	C/C	A/A
*FY*A/FY*B*	FY008.50	Fy (a+b+)	A/A	G/G	C/C	C/C	T/T	C/C	T/T	C/T	DEL/DEL	A/A	G/A	C/C	G/A	G/A	C/T	A/G
*FY*A/FY*B*	FY008.51	Fy (a+b+)	A/A	G/G	C/C	C/C	T/T	C/C	T/T	T/T	DEL/DEL	A/A	G/A	C/C	G/A	G/G	C/C	A/G
*FY*A/FY*B*	FY008.52	Fy (a+b+)	A/A	G/G	C/C	C/C	T/T	C/C	T/T	C/T	T/DEL	A/A	G/A	C/C	G/G	G/G	C/C	A/G
*FY*A/FY*B*	FY008.53	Fy (a+b+)	A/G	G/G	C/T	C/T	T/T	C/C	C/T	C/C	T/DEL	A/A	G/A	C/C	G/G	G/G	C/C	A/A

3.4 Discussion

3.4.1 Next Generation Sequencing of the *FY* library

For the construction of *FY* libraries for our NGS genotyping protocol, several optimisation steps were applied to primers, purification and size-selection so as to achieve an effective protocol. SPRIselect® reagent kit (Beckman Coulter, UK) was preferred for purification and size selection over the Pippin Prep™ instrument due to its advantages of being less time consuming and more cost-effective. In addition, the Pippin Prep™ instrument was limited in the number of samples per cassette and required an extra step after size selection (purification), thereby increasing workload and the error rate (section 2.2.4.10). *FY* templates were loaded onto a 316™ chip, despite the fact that a 314 ™ chip would have been sufficient for our *FY* samples and is cheaper. The reason for using the 316™ chip here was due to an initial unsuccessful loading of samples onto a 314 ™ chip, which resulted in a poor loading density (data not shown). This issue was mentioned in the manufacturer technical support (Ion Technical Support, 2013) and was explained by its smaller flow cell, which hindered the movement of loading liquid across the chip surface.

3.4.2 Quality control of Next Generation Sequencing data

The quality check of the high-throughput NGS data gives an initial impression of the validity of the data for downstream analysis. FastQC, which can be either used as a plugin for Torrent Suite software or independently, provides a quality report of both the sequencer instrument (here this was the Ion PGM™) and the library (Andrews, 2016). According to the Phred score, the quality of the NGS data was above 99% base call accuracy, with a 1 in 1000 probability of an incorrect base call. This high accuracy, along with a good mean coverage depth of 5600X, increases the confidence of the data

for further analysis, such as variant analysis. With regard to the accepted coverage depth in genotyping, there is as yet no fixed guideline for the number of times the target gene should be sequenced. There are various views on this issue; for example, in a study carried out by the Bentley group in 2008, in which the human genome was sequenced by massively parallel sequencing technology (using an Illumina platform), an average of 40X read depth was determined to cover the whole human genome accurately, with all homozygous and heterozygous polymorphisms detected at 15X and 33X, respectively (Bentley et al., 2008). Using a similar approach, Wang's group set a standard of 36X for their study, in which whole genome sequencing of an Asian individual was first described and provided high accuracy sequencing reads (Wang et al., 2008), while another group used 30X of depth to sequence the first Korean genome (Ahn et al., 2009). In 2011, in a study by Ajay and colleagues, whole genome sequencing (WGS) was conducted using high throughput sequencing technology and aimed to establish a sequencing guide, in terms of coverage, that reflects accurate and reliable polymorphism calling. The group found that not only was 50X adequate coverage for accurate base calling for WGS, but 35X coverage depth actually yielded comparable accuracy with updated software and improved newer sequencing chemistry (with reduced GC bias) (Ajay et al., 2011). Both Stabentheiner (2011) and Fichou (2014) agreed on the coverage depth of 50X for identification of heterozygous polymorphisms when genotyping the *RHD* gene in the former study, and sufficient to genotype 18 genes involved in 15 blood group systems using the Ion AmpliSeq™ Library on the Ion PGM™ in the latter study (Stabentheiner et al., 2011, Fichou et al., 2014). Furthermore, it is speculated that a coverage of at least 30X may be sufficient (Tilley and Grimsley, 2014) as was used in a study by Lane's group (2016), in which WGS data of at least 30X was used for studies to predict red blood cell and platelet antigens (Lane et al., 2016). Despite all these views, the coverage depth for sequencing *FY* – which can be

defined as the number of times that the reads and bases are sequenced and aligned to the reference sequence, thereby conveying the confidence level of the variant (Sims et al., 2014) – was significantly higher than that in other reports of NGS-based genotyping of human blood groups, such as that of Stabentheiner's group (2011). The reason for this considerable number is suggested to be that only a fraction of the 316™ chip capacity was used in each run. With regard to the scalability and the number of samples that can be sequenced for *FY* genotyping in a single run, different aspects are considered, such as the coverage depth, the read length and the target size. Applying the calculation (mentioned in Chapter 1), using the 200 base reads, up to 83 samples can be simultaneously sequenced for the entire *FY* gene and flanking regions when using the 314™ chip (version 1), if the desired coverage depth was 50X, as was suggested by Stabentheiner et al. (2011) and Fichou et al. (2014). This chip (314™) has the lowest capacity among other sequencing chips. Similarly, more than 800 and 4000 samples can be sequenced in parallel using the 316™ and 318™ chips, respectively. These chips, however, were first versions of chips produced by Life Technologies, who have since developed newer chips (v2) and those for the Ion Proton platform, with a much higher capacity (Chapter 1). These chips allow parallel sequencing of a very large number of samples, thereby reducing the cost of sequencing each sample for all mutations, including SNPs, deletions and insertions. The approximate estimation for the cost per sample library preparation, including the PCR reaction, fragmentation, purification and the quality check by Bioanalyzer was calculated to be about £50. On the other hand, the cost of the sequencing run, using the 316™ chip, was approximately £454 – which includes template preparation and sequencing by the Ion PGM™ machine. By applying these metrics, the cost of sequencing the entire *FY* gene per sample was estimated to be only £0.6, which would be further lower with increased chip capacity. The cost of a sequence run on this platform (Ion PGM™) is suggested to be cheaper than others, such

as Illumina GAIIX. In 2013, Rieneck and colleagues conducted a study, using the latter platform, to analyse the Kell blood group K/k SNP with the purpose of predicting the antenatal phenotype of the foetal Kell blood group (Rieneck et al., 2013). The cost of reagents for a single sequence was approximately $1500 (£1039.61), which is more than double the cost of the platform used here. Thus, it appears to be more cost-effective to perform more sequence runs with Ion PGM^{TM}.

3.4.3 Validity of *FY* Next Generation Sequencing genotyping

Consequently, having evaluated the sequencing data from *FY* samples, no further validation approaches, such as Sanger sequencing, were required for several reasons. As the linear relation of coverage depth and base calling accuracy was previously agreed, in several studies (Ajay et al., 2011, Lane et al., 2016), the coverage here was considerably high, in addition to a base call accuracy of nearly 99.99% – a Phred score of 30, which was higher here than in Ajay's 2011 study (Phred score of 20+). The critical *FY*A/FY*B* SNP was in complete concordance with the Fy$^{a/b}$ serology. Moreover, all the mutations found during *FY* sequencing did not raise any further investigations as they are common and had already been reported in the dbSNP NCBI database (and those polymorphisms were marked as validated by previous groups).

3.4.4 *FY* Next Generation Sequencing genotyping

FY genotyping is suggested to be highly beneficial by enabling more comprehensive understanding of the molecular mechanisms underlying the alleles, as well as their correlation with the phenotypes provided by serology. Castilho (2007) suggested that genotyping can provide a higher inventory of blood units for transfusion-dependent (SCD) individuals with the Fy (a-b-) phenotype, which manifests from a SNP in the

GATA promoter. In the study, 28 individuals (with a Fy (a-b-) phenotype) received Fy b+ blood without any signs of anti-Fy^b production, which could be due to the expression of the antigen on non-erythroid cells (Castilho, 2007). From these individuals, 3 were heterozygous for the *FY*B* GATA SNP, while 25 were homozygous. High throughput genotyping platforms, based on microarray technology, have illustrated better interpretation and understanding of the molecular mechanisms of the *FY* phenotype; for example, Hashmi and colleagues (2005) have shown that, by using HEA BeadChip, 1 out of 52 patients phenotyped as Fy b- lack the *FY*B* allele, while the rest carry that allele, but contains the GATA SNP (Hashmi et al., 2005). As a result, according to the assumption of Castilho, they (51) could receive Fy b+ blood components. Despite the advantages of microarray-based platforms, such technologies have the disadvantage of not being able to identify novel mutations that could affect antigenicity; therefore, NGS is suggested to address this issue.

To our knowledge, this is the first NGS-based genotyping study of the *FY* gene using LR-PCR, in which comprehensive genotyping and analysis of polymorphisms of the entire *FY* gene (intron, exons and outer regions) was carried out. Some crucial aspects were taken into consideration, such the ability to couple NGS with LR-PCR rather than another approach, such as exome sequencing, where mostly coding regions (exons) are targeted. Sequencing of the whole gene enables the analysis of polymorphisms in, exons, introns and outer regions; for example, identification of the silencing SNP (- 67 T>C) in the promoter region resulted in higher resolution, complete genotyping and, thus, better association with the phenotype.

3.4.4.1 Mutations in exons and the promoter region

43 out of 53 samples were of known serological phenotype with regard to the Fy^a/Fy^b antigens. The NGS genotyping data matched and confirmed the phenotypes of *FY* alleles *FY*A*, *FY*B* and *FY*02*(*Null*) in terms of the crucial SNPs previously described (Reid et al., 2012), while the SNP and amino acid locations were based on the major *FY* form (Daniels, 2013). The crucial SNP *FY*A*/*FY*B* (125G>A) in exon 2, encoding for Gly42Asp, was in concordance with the Fy^a/Fy^b phenotype, as was the zygosity. The allele *FY*02N.01* was found in 7 out of 53 samples and results from the aforementioned silencing SNP (- 67 T>C) in the promoter region, causing the binding of the erythroid transcription factor (GATA1) to be disrupted. This allele was found to be homozygous in all 7 samples and impaired $Fy^{a/b}$ antigenicity of the erythrocytes, which manifested with the phenotype Fy (a-b-) samples carrying a *FY*B* background with homozygous 125G>A. The missense SNP 298G>A in exon 2 is common and was found in 16 out the 53 samples, all of which expressed the *FY*B* allele (17 *FY*B* haplotype); none of the samples genotyped as *FY*A*, *FY*A* (16/53) showed the SNP 298A. This SNP leads to the amino acid change Ala100Thr, which has been found to be expressed along with another polymorphism – SNP 265 C>T in exon 2 (amino acid change Arg89Cys) of the *FY*X (FY*02M.01)* allele (Olsson et al., 1998, Yazdanbakhsh et al., 2000). This allele encodes reduced expression of Fy^b antigens; however, the existence of 298G>A (Ala100Thr) alone may not affect the expression of Fy^b antigens, but rather may play a cumulative role with SNP 265 C>T, as suggested by (Gassner et al., 2000) and Olsson et al. (1998). The latter author found this SNP in 22% of samples with the normal Fy (a-b+) phenotype (Olsson et al., 1998) and the frequency of this SNP alone is near to that found in Caucasians with *FY*B* (around 15%) (Reid et al., 2012) . The *FY*X (FY*02M.01)* allele encodes for weakened expression of Fy^b antigens, as shown by the weak reaction of these samples with anti-Fy^b (Olsson et al., 1998, Gassner et al., 2000),

and was found here in 1 sample; however, despite the presence of that allele, the sample showed normal Fy^b antigenicity with a Fy (a-b+) phenotype. Nevertheless, it might be argued that this sample was Fy^b+ due to both SNPs (265 C>T and 298G>A) being heterozygous; however, weak reactivity was found in 2 patients, one of whom carried *FY*B*, *FY*X* and the other was homozygous for both SNPs (265 C>T and 298G>A) (Olsson et al., 1998, Daniels, 2013). The NGS approach was able to show that this weakening allele (*FY*X*), in contrast to serology, may not be detected in samples initially typed as Fy (a+b-) but with *FY*A* and *FY*B* (with weak anti-Fy^b serological reaction), which is suggested to be due to the *FY*X* allele (Murphy et al., 1997).

Another mutation in exon 2 was a 714G>A SNP that has no apparent effect on the amino acid (Gly283Gly). As a result, the $Fy^{a/b}$ antigenicity was not affected due to normal expression and a phenotype of Fy (a+b+); thus, according to the *FY*A, FY*B* SNP (125G>A), the genotype was *FY*A/FY*B*. The observed low frequency of this SNP here (1 sample in 53), was in agreement with the data provided by the 1000 genomes project (above 2% in the UK, 1% in Europe and 0.3% worldwide).

3.4.4.2 Mutations in introns

Due to the extensive data that can be obtained from NGS genotyping, analysis of mutations, such SNPs and deletions, in introns (which be allele specific) of the *FY* alleles could enable refining of the reference sequence for those. However, despite finding mutations in intronic regions of this gene, there was no complete correlation observed with the *FY*A*, *FY*B* mutations. A high frequency deletion (77% in Europe, 1000 genomes; T>Del 159175005) was found in samples and, although it was shown to

be associated with a *FY*B* allele background, it was also found in 14 *FY*A* haplotypes. All 53 samples (106 haplotypes) were homozygous for the G>C 159174824 SNP, which implies that all our sequenced samples deferred from the reference sequence derived from the Human Genome Project (hg19). One explanation for this could be that this sequence variation is rare but found when the human genome was assembled.

The SNP catalogue approach, collated here (Table 3.7), using a colouring scheme, lists the mutations provided by NGS across the predicted genotype and phenotype, allowing simplified analysis of different allele patterns. In 2000, Gassner and colleagues suggested that the *FY*X* allele may arise from 3 point mutations: 265 C>T (Arg89Cys) and 298G>A (Ala100Thr) in exon 2 and an intronic SNP (190 C>T, where the first T in the intron is number 1). The latter SNPs may exert a cumulative effect with the former in reducing Fy^b expression (Gassner et al., 2000). Although the chromosomal location was not mentioned, it could be SNP 159174920 (C>T) (as all intronic SNPs of *FY* in our samples were analysed, see Table 3.7). From Table 3.7, it can be seen that in samples of *FY*B* background, 298G>A (Ala100Thr) persistently coexists with 159174920 T. Although there was no clear reference intronic SNPs (allele-specific SNPs), different *FY* alleles could be pointed out by looking at the various patterns in Table 3.7; for example, 4 *FY*A* alleles, at least 4 *FY*B* and 2 *FY*02N.01*. These intronic SNPs are catalogued, although their effect on the antigenicity is not yet known; nevertheless, they may be useful to consider when designing primers to avoid primer binding site issues. The new alleles with different patterns of intronic SNPs may be further studied in a wider population to assign the frequency of alleles to different ethnic backgrounds. This was in fact accomplished by a study by Schmid et al (2012), in which the frequency of *FY* alleles among 54 African American was investigated by high resulution genotyping that included polymorphism analysis in the exons, intron and untranslated regions. According to the intronic polymorphisms (those highlited in

Table 3.6), along with those in exons and the GATA box, 11 diiferent *FY* alleles were described, including 3 *FY*A,* 4 *FY*B,* 2 *FY*02N.01,* 1 *FY*B* with 298G>A and 1 *FY*02 M.01*. In addition, the author suggested terminology to reflect the high resolution genotyping, for example for the 3 different *FY*A* alleles (*FY*01:1, FY*01:2 and FY*01:3*). Moreover, it was suggested that the extensive genotyping, especially the polymorphisms in introns, may help in the studies of evolution (Schmid et al., 2012).

3.4.5 The advantage of Next Generation Sequencing genotyping of *FY*

Here, the NGS shows superiority over other high throughput genotyping approaches (for example, microarray platforms) in that all existing mutations within the *FY* gene in all 53 samples were revealed, including introns and flanking regions. This comprehensive method of genotyping could provide a better picture of the phenotype from the genotype. In addition, NGS is capable of discovering novel SNPs (due to genotyping be sequencing) that may account for new antigenicity-affecting alleles, in contrast to SNP probe-reliant platforms that fail to detect novel SNPs and may falsely predict a positive phenotype encoded by rare silencing or weak alleles. To prevent this, the probe list needs to be updated frequently to prevent negative consequences, such as allelic dropout of low frequency alleles (McBean et al., 2014, Avent et al., 2015, Tilley and Grimsley, 2014).

Examples of new *FY*B* alleles that silence Fy^b expression (Fy b-), were described in 2014 by Westoff's group, in 2 samples (Westoff et al., 2014). The phenotype of the first sample was Fy (a-b-) but was predicted as Fy (a-b+) by HEA BeadChip. A sequencing approach then revealed a 2 nucleotide deletion, which was suggested to lead to a frameshift and formation of a premature stop codon, 179_180delCT (Ser60CysfsTer16) in exon 2 (Westoff et al., 2014). Similarly, the second sample was in concordance with the Fy^b phenotype, as suggested by DNA testing; however, sequencing revealed a

missense mutation in exon 2 895G>A (Ala299Thr) that was thought to be a silencing SNP. The frequency of these alleles were predicted to be low, thus, corresponding probes were not added to the HEA BeadChip assay, although they will presumably need to be added subsequently (Westoff et al., 2014).

Furthermore, it was pointed out that the commercially available microarray high throughput platforms, such as BloodChip and BeadChip, were designed to respectively cover SNPs of specific populations, namely European and American (including African-American). This is advised to be considered when using such platforms in a multi-ethnic population, such as Australia (McBean et al., 2014). NGS, on the other hand, circumvents this issue due to its sequencing-based approach which provides discovery capability (Avent et al., 2015).

In conclusion, NGS with LR-PCR proved to be capable of genotyping the *FY* gene extensively. Using only one single PCR product, all existing mutations (in introns, exons and regulatory regions) were identified. Moreover, due to its high throughput sequencing capability, NGS features a discovery mode that can reveal novel SNPs and alleles in a large number of samples, allowing more accurate prediction of antigen phenotypes. This surpasses the microarray-based method, in which predefined knowledge of SNPs is required, with rare or novel SNPs frequently missing from the probe list and thus needing further investigations.

The high throughput productivity of NGS allows a significant number of samples to be sequenced at a high resolution for all mutations in a single run, thereby reducing costs. In addition, it allows a comprehensive scan of intronic SNPs, which can be further studied for association with *FY* alleles and considered for primer design. Importantly, NGS allows for more comprehensive association analysis between the genotype and the phenotype. In addition, if applied to a large cohort, reference sequences for *FY* alleles

could be provided and the prevalence of different polymorphisms among different populations can be assessed. Furthermore, applying this approach of extensive NGS genotyping, analysis and cataloguing of SNPs, cases with unusual discrepancies between genotype and phenotype can be investigated, which ultimately enables the direction of the interpretation from genotype to phenotype.

Finally, this chapter has illustrated that this NGS protocol can be applied to genotyping *FY* and, probably, for other more complex blood groups, such as *JK* and *ABO* (discussed in the following chapters) – and thus is potentially important towards obtaining a more comprehensive profile of blood group mutations and, thus, facilitating the provision of compatible blood that will improve transfusion safety.

Chapter 4

Genotyping of the JK Blood Group by Next Generation Sequencing

4.1 Introduction

The Kidd blood group system (JK/ISBT 009) is represented by a single gene (*JK* or *SLC14A1*) (see section 1.7.1) comprising 11 exons, with the translation start codon located in exon 4, encoding for 389 amino acids (Lucien et al., 1998, Horn et al., 2012). The main polymorphic Jk antigens, $Jk^{a/b}$, are thought to arise from a single SNP (838G>A in exon 9) within the co-dominant *JK*A/JK*B* alleles, encoding amino acid change Asp280Asn (Olivès et al., 1997). According to The NCBI Blood Group Antigen Gene Mutation Database (BGMUT), additional *JK* alleles (36 in total) have been uncovered due to many additional polymorphisms (dbRBC, 2016), some of which might reduce or remove expression of the Jk antigens (Reid et al., 2012), with the number of *JK* alleles continually rising due to novel polymorphisms (Keller et al., 2014). These polymorphisms may cause discrepancy and false interpretation of the Jk phenotype and antigenicity, which carries the risk of adverse reactions during clinical blood transfusions. There is evidence of both $Jk^{a/b}$ antibodies causing delayed haemolytic transfusion reactions (DHTRs) (Hussain et al., 2007, Pineda et al., 1999) and haemolytic disease of the foetus and newborn (HDFN), although the latter was less common (Ferrando M, 2008, Daniels, 2013). There is, therefore importance in expanding genotyping of the clinically-significant blood group systems (including genes such as *JK*) as a means of ensuring the safety of blood transfusions and reducing alloimmunisation – especially in patients requiring multiple blood transfusion (for example, those with sickle cell disease) (Avent et al., 2015). A high-throughput

genotyping approach has been suggested to tackle the large demand for screening the high number of alleles (Anstee, 2009). However, all commercially available platforms for high-throughput genotyping, particularly microarray technology, currently rely on known polymorphisms and alleles; thus, they are unable to detect new or rare polymorphisms not already included in the array and need constant updating as a result. Next Generation Sequencing (NGS) is a recently introduced high-throughput sequence-based genotyping technology that is capable of detecting new polymorphisms, thereby enabling extensive genotyping of blood group genes and accurate interpretation of the associated phenotypes.

4.2 Aim of the study

Similarly to the *FY* chapter (section 3.2), the aim of this study was to combine the powerful capability of NGS, together with long-range polymerase chain reaction (LR-PCR) to establish a protocol to comprehensively genotype and analyse the entire *JK* gene. The discovery mode of NGS will allow detection of all existing, rare or new polymorphisms of both intronic and exonic origin. The correlation of all detected polymorphisms was assessed with key *JK* SNPs in order to evaluate their allele specificity and uncover any allele-specific sequences (‘fingerprints’) that may also be associated with weakening or silencing *JK* alleles. This would also allow more accurate analysis of the molecular basis of *JK* alleles, thus, better prediction of the resulting phenotype. The feasibility of using NGS for genotyping of *JK* and other blood groups, such as ABO, was also discussed here.

4.3 Results

4.3.1 LR-PCR of the *JK* gene

The gDNA of 67 samples was used for extensive genotyping of *JK* by NGS. Serology information on the Jk (a/ b) phenotype was provided for most of the samples (59/67) and was used for sample selection (Table 4.1). Using Primer 3 software and the NCBI database (see section 2.2.4.1), a total of 3 primer pairs were designed to produce 3 overlapping amplicons of various sizes (11012, 11053 and 14665bp, see Table 2.2). The entire *JK* gene plus flanking regions was amplified, as illustrated in Figure 4.1. For the LR-PCR amplification reaction, LongAmp® Hot Start Taq Master Mix was used under thermocycling conditions (Table 2.5; section 2.2.4.2). The first two amplicons (11012bp and 11053bp) were loaded onto 1% (w/v) agarose gel and appeared as single bands above 10kb (Figure 4.2). The third amplicon (14665bp) was loaded onto an 0.8% (w/v) agarose gel and appeared as a single band below 20kb (Figure 4.2; section 2.2.4.3). The amplicons were then purified using a magnetic bead technique (section 2.2.4.4) and quantified using the Qubit® 2.0 Fluorometer with the broad range (BR) assay kit. This enabled preparation of the required (100 ng) of amplicons required for the fragmentation process. To ensure equal representation of the 3 *JK* amplicons, approximately 33.3ng (100/3= 33.3) of each amplicon was pooled for the fragmentation reaction (section 2.2.4.5).

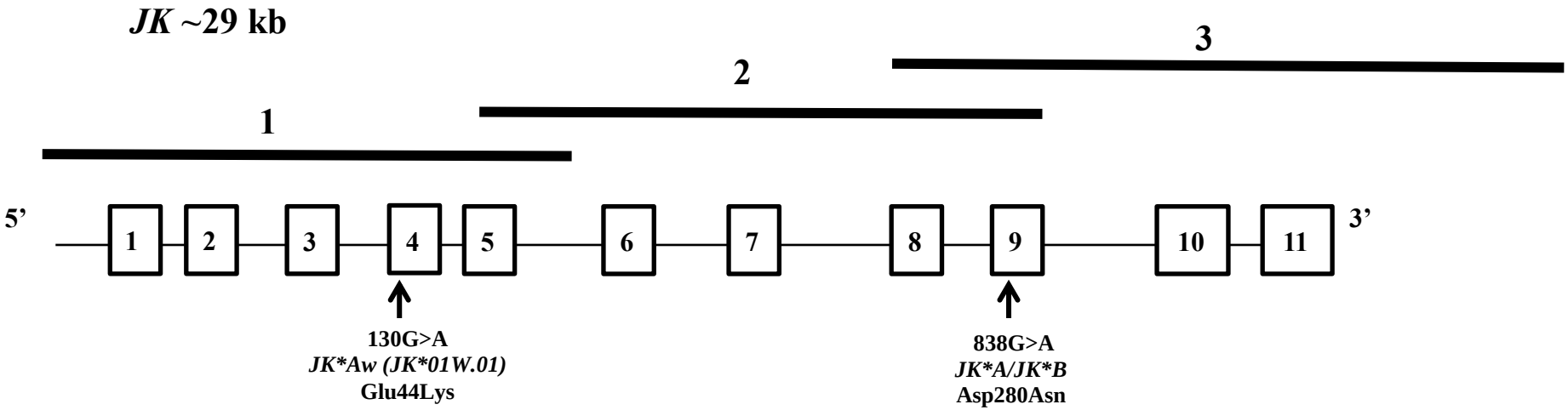


Figure 4.1 The *JK* gene is illustrated with the three overlapping long amplicons generated by LR-PCR: 1 (11012bp), 2 (11053bp) and 3 (14665bp). These amplicons cover the entire *JK* gene and flanking regions.

The *JK* gene is 28765 bp in length (according to NCBI GenBank), comprising 11 exons (shown as boxes) and 10 introns (represented by a fine line). The crucial SNP that generates *JK*A* and *JK*B* is located in exon 9, while the suggested SNP responsible for generation of the *JK*01W* allele (*JK*01W.01*) is located in exon 4. Refer to Table 2.2 for more details on amplicon location.

Table 4.1 Serology information on the 67 blood samples provided by the National Health Service Blood and Transplant (NHSBT; Filton, Bristol UK). Asterisks (*) denotes samples for which sample ID were provided. ND, serology not defined. **JK009** is the ISBT number assigned for the JK blood group system (Daniels, 2013, Reid et al., 2012) and is used here to name samples in accordance with the *JK* NGS genotyping experiment.

Sample number	Sample ID*	Phenotype	Sample number	Sample ID*	Phenotype	Sample number	Sample ID*	Phenotype
JK009.01	RJ	Jk(a+b-)	JK009.26	39	Jk(a+b-)	JK009.51	54	Jk(a-b+)
JK009.02	4	Jk(a+b-)	JK009.27	R2	Jk(a+b+)	JK009.52	55	Jk(a-b+)
JK009.03	33	Jk(a+b-)	JK009.28	4715	ND	JK009.53	56	Jk(a-b+)
JK009.04	580X	ND	JK009.29	RG	Jk(a-b+)	JK009.54**	470P	(Jka-b+)
JK009.05	RD	Jk(a+b-)	JK009.30	RN	Jk(a-b+)	JK009.55*	RH	Jk(a+b+)
JK009.06	6	Jk(a+b-)	JK009.31	RX	Jk(a-b+)	JK009.56	R9	Jk(a+b+)
JK009.07	16	Jk(a+b-)	JK009.32	RK	Jk(a-b+)	JK009.57	RB	Jk(a+b+)
JK009.08	13	Jk(a+b-)	JK009.33	R7	Jk(a-b+)	JK009.58*	RL	Jk(a+b+)
JK009.09	32	Jk(a+b-)	JK009.34	R8	Jk(a-b+)	JK009.59*	R97	Jk(a+b+)
JK009.10	35	Jk(a+b-)	JK009.35	48	Jk(a-b+)	JK009.60*	2	Jk(a+b+)
JK009.11	36	Jk(a+b-)	JK009.36	49	Jk(a-b+)	JK009.61	17	Jk(a+b+)
JK009.12	37	Jk(a+b-)	JK009.37	5800	ND	JK009.62	453R	ND
JK009.13	28	Jk(a+b-)	JK009.38	437R	Jk(a-b+)	JK009.63	469B	ND
JK009.14	40	Jk(a+b-)	JK009.39*	3	Jk(a-b+)	JK009.64	4208	Jk(a+b+)
JK009.15	438P	Jk(a+b-)	JK009.40	7	Jk(a-b+)	JK009.65	4576	Jk(a+b+)
JK009.16	79	Jk(a+b-)	JK009.41	8	Jk(a-b+)	JK009.66	469Z	ND
JK009.17	82	Jk(a+b-)	JK009.42	22	Jk(a-b+)	JK009.67	476D	ND
JK009.18	86	Jk(a+b-)	JK009.43	42	Jk(a-b+)			
JK009.19	RP	Jk(a+b-)	JK009.44	43	Jk(a-b+)			
JK009.20	4680	ND	JK009.45	44	Jk(a-b+)			
JK009.21	5	Jk(a+b-)	JK009.46	45	Jk(a-b+)			
JK009.22	41	Jk(a+b-)	JK009.47	46	Jk(a-b+)			
JK009.23	26	Jk(a+b-)	JK009.48	47	Jk(a-b+)			
JK009.24	579J	Jk(a+b-)	JK009.49	50	Jk(a-b+)			
JK009.25	81	Jk(a+b-)	JK009.50	51	Jk(a-b+)			

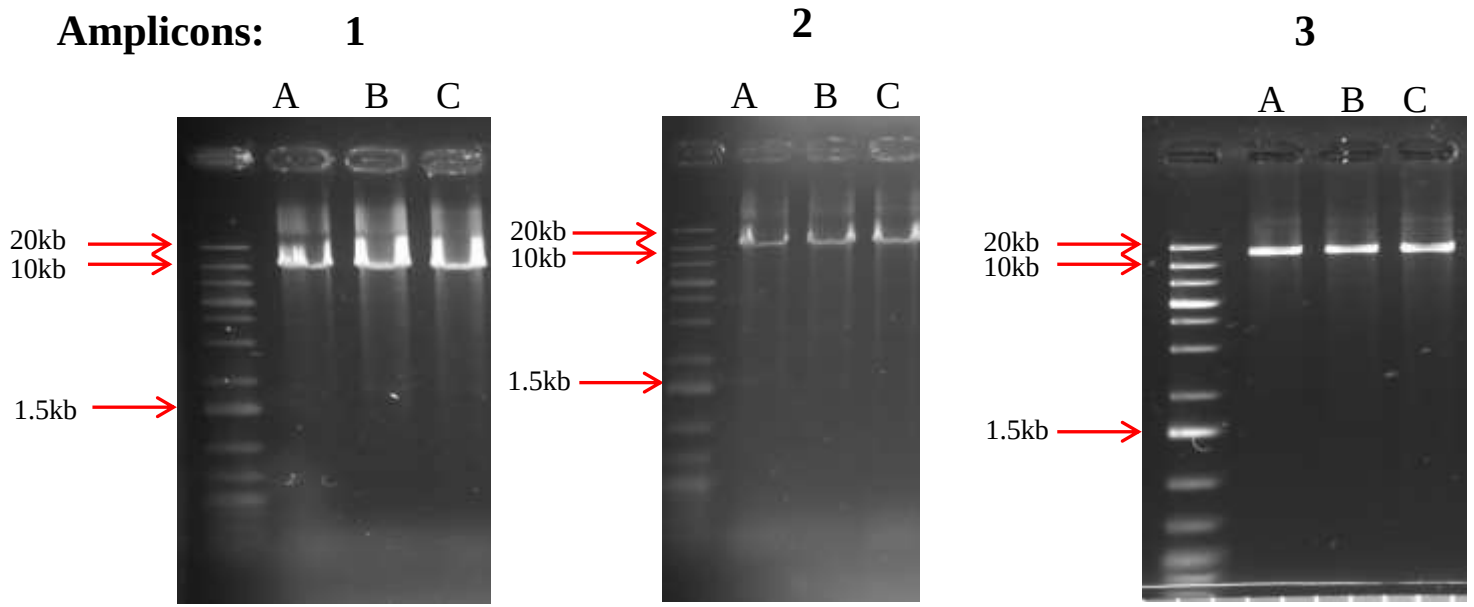


Figure 4.2 Amplification of the entire *JK* gene using 3 amplicons produced by LR-PCR.

For each sample, the entire *JK* gene and flanking regions were covered by 3 amplicons of different sizes: 1 (11012bp), 2 (11053bp) and 3 (14665bp). The above images are examples of amplicons from three samples, A (JK009.04), B (JK009.62) and C (JK009.67). Amplicons 1 and 2 were loaded onto 1% agarose gel, while amplicon 3 was loaded onto 0.8% agarose gel due to large size (~15kb) to allow easier migration of the band. All amplicons were electrophoresed at 70 V for 1 hour and 20 minutes. The GeneRuler^{TM}1Kb Plus DNA ladder (Thermo Fisher Scientific) was used here as a marker of DNA size.

4.3.2 NGS the *JK* gene

4.3.2.1 *JK* amplicon library fragmentation (purified-fragmented library)

Fragmentation of NGS libraries was conducted following amplicon purification and quantification. For each sample, the 3 amplicon-pool (100ng) was fragmented using Ion Xpress^{TM} Plus Fragment Library Kit (section 2.2.4.6). The enzymatic fragmentation reaction was incubated for 8 minutes, after which samples were purified using magnetic beads (section 2.2.4.6). The Agilent® 2100 Bioanalyzer and Agilent High Sensitivity DNA Kit (section 2.2.4.7) were used to assess fragment size distribution. Figure 4.3 depicts the size distribution of a fragmented-purified *JK* sequencing library.

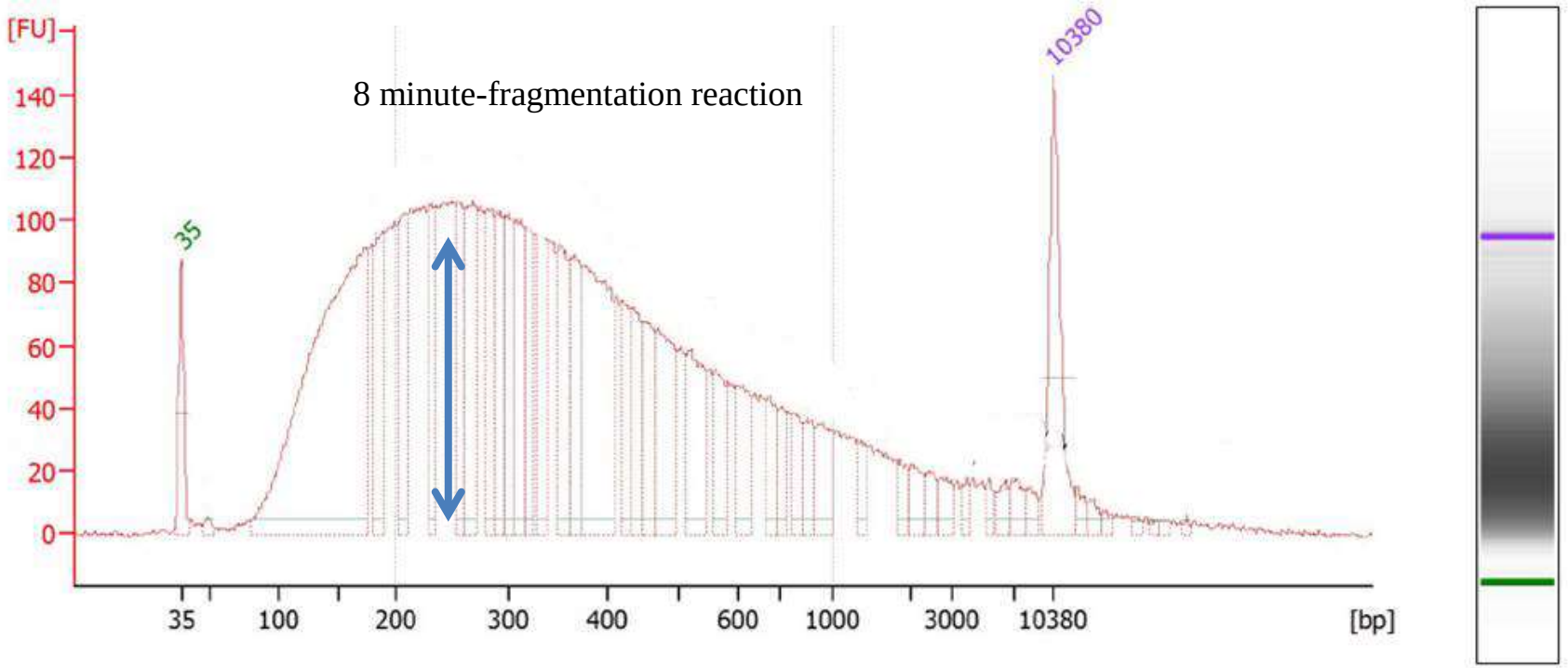


Figure 4.3 An electropherogram of a fragmented-purified *JK* DNA library (consists of a pool of 3 amplicons).

An example of a *JK* sample fragmented using the Ion Xpress™ Plus Fragment Library Kit for 8 minutes. A wide fragment size distribution can be seen with a peak around 200-300bp (marked by a blue arrow), which is recommended for the 200bp read length on the Ion PGM™. The green line (35bp) is the lower marker and the purple line (10380bp) is the upper marker. Results shown were obtained using the Bioanalyzer® 2100 instrument.

4.3.2.2 Ligation of barcoded adapters (*JK* purified-ligated library)

The fragmented-purified samples were ligated to barcoded adapters, in which P1 and the Ion Xpress™ Barcode X adapter (provided by the Ion Xpress™ Barcode adapters Kit) were used along with DNA ligase. The ligated libraries were then purified and analysed using Agilent® 2100 Bioanalyzer with Agilent High Sensitivity DNA Kit (section 2.2.4.9). Figure 4.4 shows the size distribution of one of the *JK* sequencing libraries after ligation.

4.3.2.3 Size selection

The ligated libraries were then size-selected to provide a suitable size range (around 200bp) for the Ion PGM™ Template OT2 200 Kit. Sequencing of the 67 samples was conducted in four separate experiments. In the first two experiments, 12 and 20 samples were sequenced and size-selected using the Pippin Prep™ instrument respectively. In the remaining experiments (17 and 18 samples respectively), samples were size-selected using SPRIselect® reagent magnetic beads. The SPRIselect® was selected over Pippin Prep™ instrument for the last two experiments due to its technical advantages (see section 2.2.4.10). Figure 4.5 depicts two size-selected *JK* samples by the two approaches. The concentrations of these libraries were obtained and calculated using the Bioanalyzer® 2100 instrument (section 2.2.4.11) for the purpose of template preparation (section 2.2.5) prior to sequencing using the Ion Torrent PGM™ (section 2.2.6).

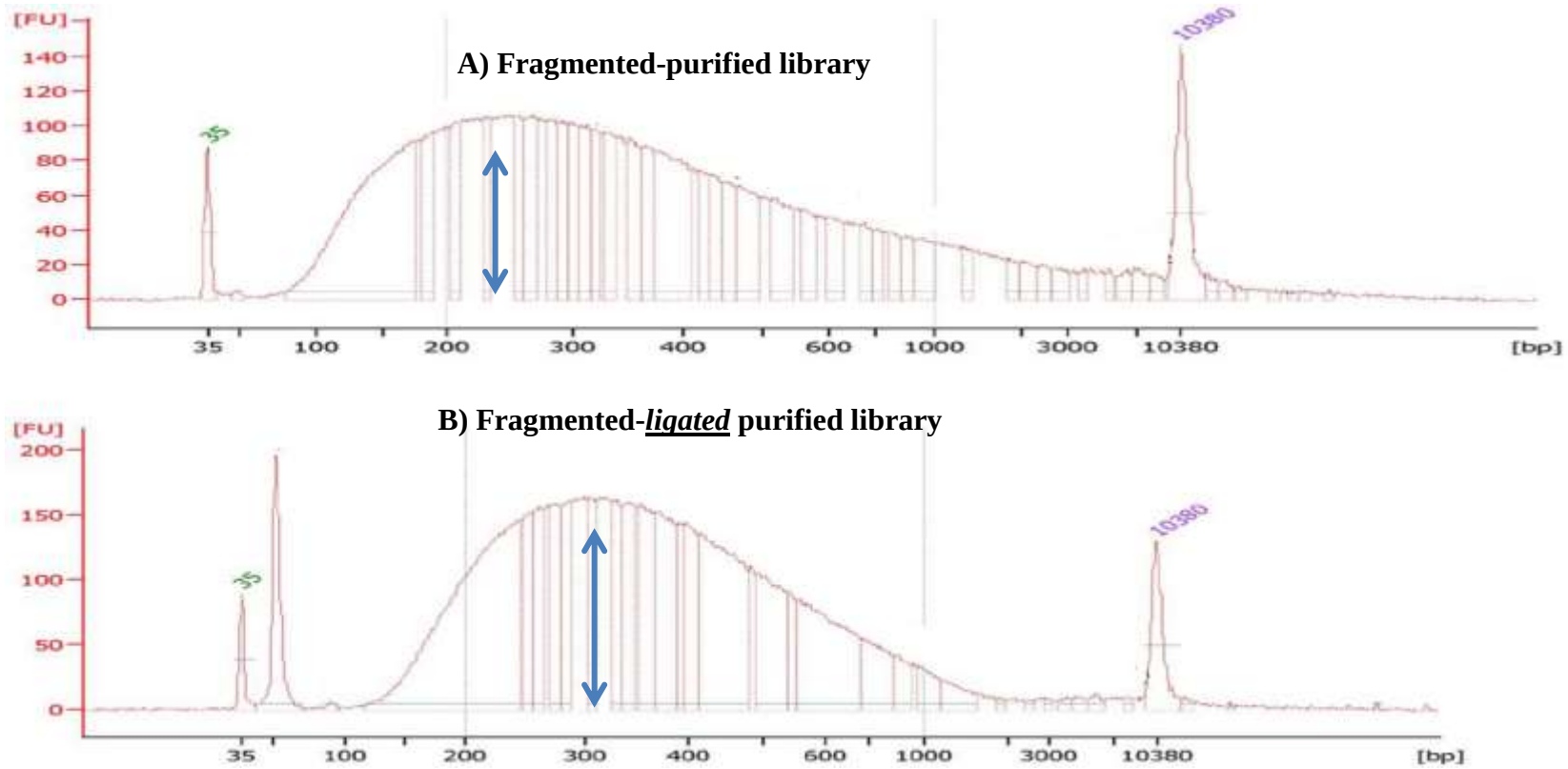


Figure 4.4 Two electropherograms of the same *JK* amplicon library.

A) A fragmented-purified library with a peak of around 200-300bp.

B) A fragmented-ligated purified library showing a peak shift to the right (caused by adapter ligation (size about 80bp), evident as an increase in size).

The green line (35bp) is the lower marker and the purple line (10380bp) is the upper marker. Results shown were obtained using the Bioanalyzer® instrument.

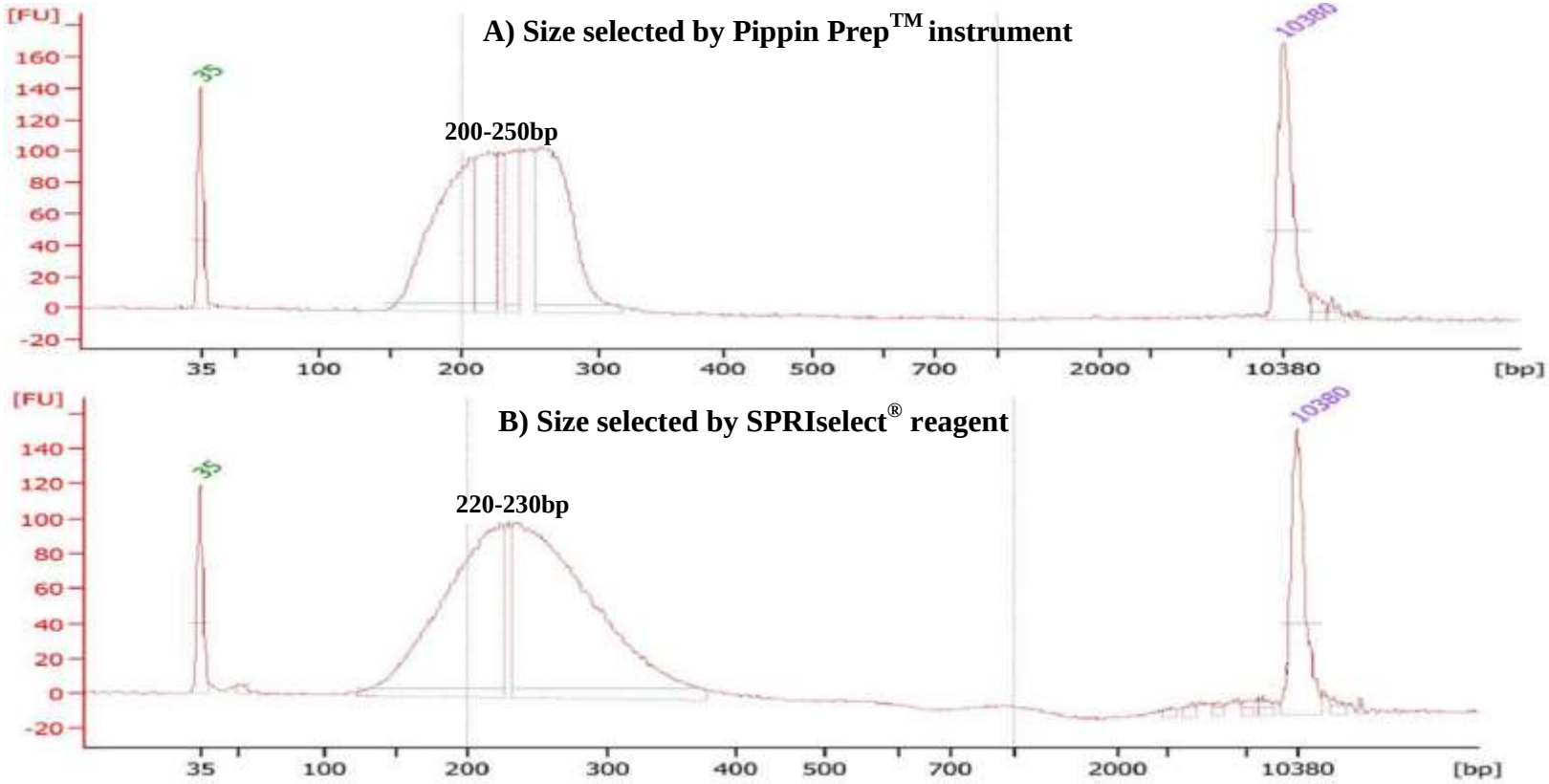


Figure 4.5 Electropherograms of two different *JK* DNA libraries size-selected using either Pippin Prep™ or SPRIselect®.

The peaks obtained from the two different approaches were comparable.

A) *JK* library size-selected by Pippin Prep™, in which a broad range was selected (around 200-250bp).

B) JK library size-selected by SPRIselect® reagent to obtain a peak around 200bp (220-230bp).

The green line (35bp) is the lower marker and the purple line (10380bp) is the upper marker. Results shown were obtained using the Bioanalyzer® 2100 instrument.

4.3.3 NGS data quality control

4.3.3.1 Sequencing data summary report

The 67 *JK* samples were sequenced using Ion PGM™ in four separate runs. The data was then processed and a summary was produced by the Torrent Suite™ Software Version 4.4 (section 3.3.3.1). The average of the total reads generated from the 67 *JK* samples was 3,548,400, with a mean coverage depth of around 750x. A summary of the *JK* sequencing report from the 4 separate runs is shown in Table 4.2. A brief sequencing report of a single representative *JK* sequencing run is displayed in Figure 4.6. In this report, the ISP loading density was 88%, while 12% of the 316™ chip wells were empty. There were around 4 million usable reads in total, which is 72% of the total reads containing the library ISPs usable for downstream analysis. These reads were thoroughly processed by well classification and read filtering to ensure their quality (section 3.3.3.1). Clonal ISPs (in which single template fragments were amplified) was made up 78 % of reads, while 22 % were polyclonal ISPs. After filtering out 1 % of test fragments and 8 % of low quality reads, the final percentage of the final library was 91% with a mean read length of 147 bp (Figure 4.6).

Table 4.2 A summary of the Ion PGM™ sequence report of the 67 *JK* samples, processed in 4 runs.* 18 samples were sequenced in 3rd run and 26 in 4th run but those repeat samples were excluded from the actual genotyping analysis

	No. of samples genotyped	ISP Loading %	Total usable reads	Usable reads%	Mean read length
1st run report	12	72	2,770,072	64	131 bp
2nd run report	20	79	3,274,431	66	134 bp
3rd run report	17+1 repeat*	82	4,147,776	80	133bp
4th run report	18+8 repeats*	88	4,001,319	72	147bp

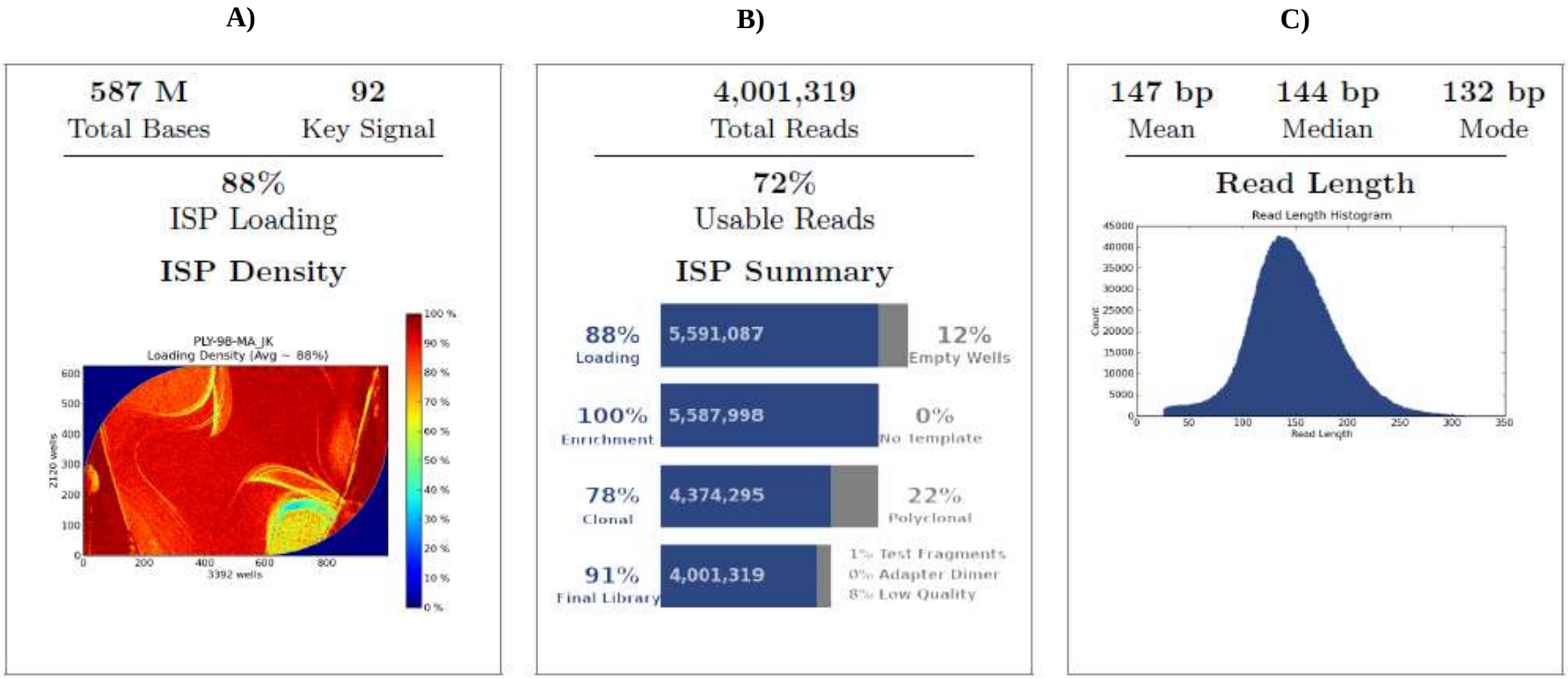


Figure 4.6 A report of a representative single sequencing run of the *JK* library

A) The percentage of ISP loading density of the ISP addressed by the 316^{TM} chip wells was 88 %. Colouring represents the loading percentage of ISP across the physical 316^{TM} chip plate surface (red is highest; blue is lowest).

B) The total number of usable reads is 4,001,319, after trimming and filtration from empty wells, non-templated and polyclonal reads. The percentage 72% is obtained by dividing these reads by the number of reads containing the library ISPs (5, 564, 617). The live/enrichment percentage is 100%, which indicates that ISPs contain a strong sequence signal from test fragment and library (templated).

C) Histogram shows a mean reading length of 147bp. The read count is displayed in the y-axis, while the read length, in bp, is shown on the x-axis.

4.3.3.2 NGS data quality control

Good quality NGS data is important for accuracy of further analysis, including genotyping. The mean coverage depth (750X) and the quality of the generated sequence data were assessed, in accordance with the Phred score, using the FastQC plugin on the Torrent Suite. Assessment of per base and per sequence quality was projected, by the FastQC, which provides an initial impression of the overall quality of the data.

4.3.3.2.1 Per base sequence quality

Figure 4.7 illustrates an overview of the quality of sequenced base pairs at their position on the reads from a single run, according to the Phred score. The mean quality score of sequenced bases in this run was 29-30, which gradually decreased with longer read length; this was suggested to be due to degradation of the sequencing chemistry near the end of the run (Andrews, 2016). According to the Phred score, the per base sequence quality of this run (30) indicated a base call accuracy of 99.9%, with a probability of 1 in 1000 of incorrect base call. The quality of all four *JK* sequencing runs was comparable (above 99% base call accuracy, according to the Phred score; Table 4.3).

4.3.3.2.2 Per sequence quality scores

FastQC also provides per sequence quality scores, which assess the quality of sequences and whether a subset of the generated sequences had poor quality. Figure 4.8 shows the mean sequence quality (30; according to the Phred score) of a single run, which matches the quality indication previously discussed. Thus, the base call accuracy of the run is high (99.9%). The high quality of the analysed parameters, including coverage depth, provide confidence in the high throughput data for further analysis. The quality of all four runs was comparable (above 99% base call accuracy, according to the Phred score; Table 4.3).

Table 4.3 Summary of the sequencing quality of the four *JK* NGS runs.

All runs achieved base call accuracy above 99%. The accuracy of the fourth run was mainly 99.9%. *The quality scoring was based on Phred score.

Run	Per base sequence quality*	Per sequencing quality score*
1st	28-29	28-29
2nd	28-30	29
3rd	29-30	29-30
4th	29-30	30
Base call accuracy	above 99%	above 99%

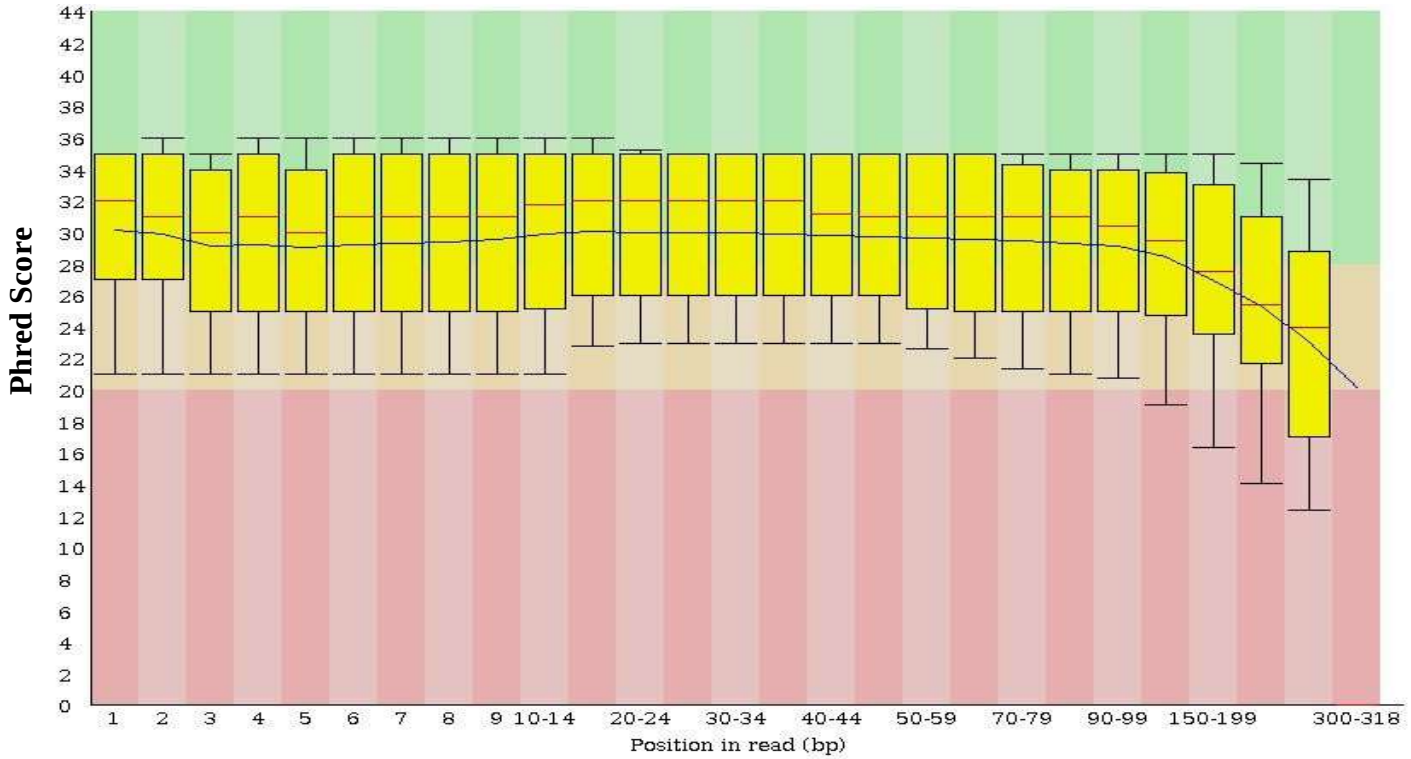


Figure 4.7 Mean Phred quality scores across all bases of *JK* samples in a single run.

The x-axis represents the position of bases in the read, while the Phred score is displayed on the y-axis. A tri-coloured background divides the y-axis into three regions of different quality levels, according to the Phred score: very good (green); reasonable (orange); and poor quality base calling (red). A Box-Whisker-type plot is drawn for each position along with a 25-75% interquartile range. Upper and lower whiskers represent the 10% and 90% points. The blue line represents the mean value of the base call quality (29-30), which mainly indicates a 99.9% base call accuracy. The median value of the quality is denoted by a red line. This plot is representative of other *JK* runs.

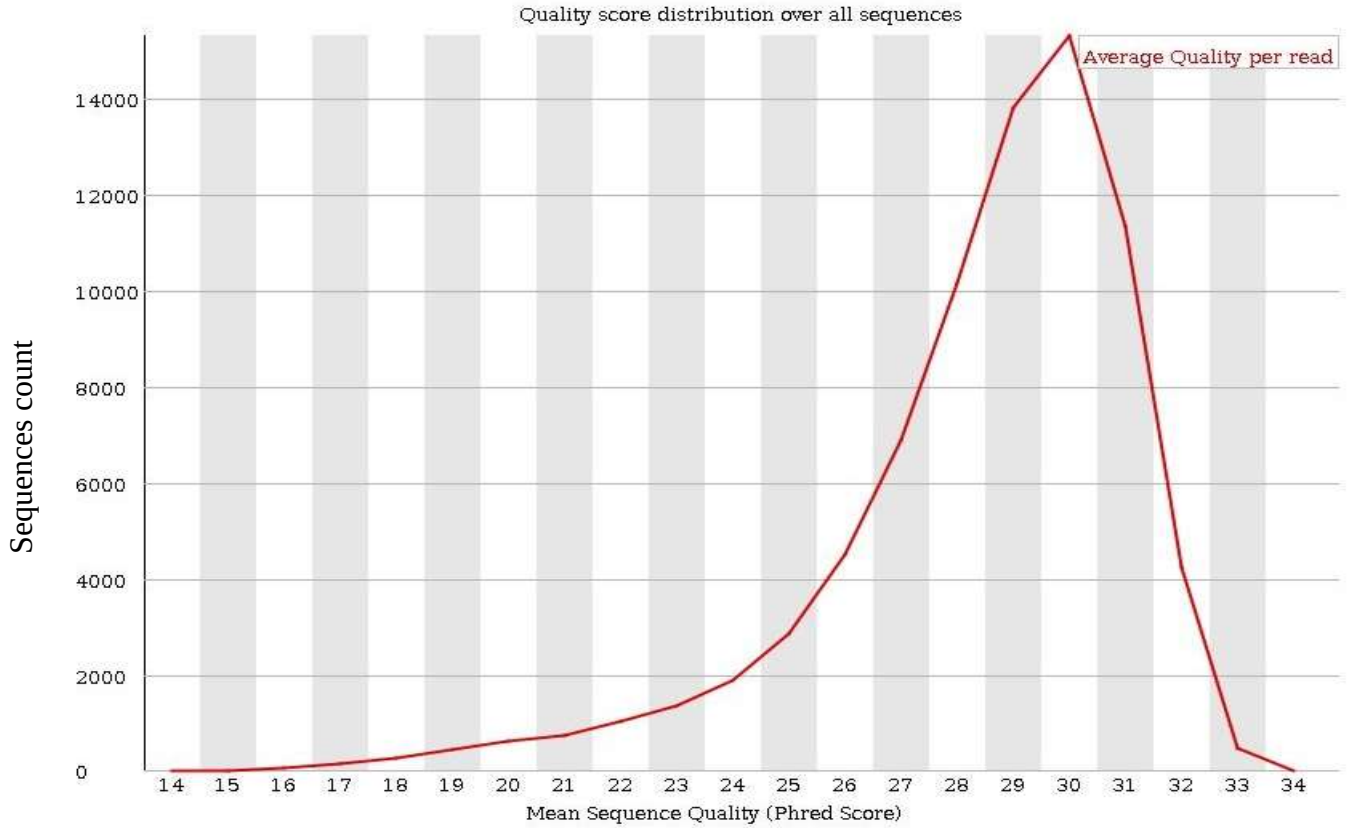


Figure 4.8 The mean quality score of the *JK* sequences generated from a single run.

The mean quality score (Phred score, x-axis) of the sequences across the number of reads (y-axis) is shown. The mean quality of most of the sequences was ~30 (Phred score), which indicates high quality of generated reads, with an accuracy of 99.9% and a probability of 1 in 1000 that the base was incorrectly called. Other runs had comparable results.

4.3.3.3 NGS sequence visualisation

Visualisation of the generated sequencing data of the *JK* samples was conducted using IGV software (2.3) and the BAM file format. The sequencing reads were aligned against the *JK* reference sequence of the human genome (hg19), associated with the accession number NM_001146036.2. The following factors were analysed using this visualisation tool: the integrity of the sequence, in terms of full coverage across the *JK* gene and flanking regions; coverage depth; and genomic information. Examples of genomic information obtained from IGV are: gene mutations, such as SNPs, insertions and deletions; zygosity; and the chromosomal location of the nucleotides. IGV was used alongside other analytical software packages, such as the Torrent Suite^{TM} plugin, Variant Caller, to aid significantly in the analysis, as previously described (Robinson et al., 2011). Figure 4.9 shows an example of this analysis, to visualise and assess a *JK* sample with serology Jk (a+b+). The integrity of the library preparation step and the specificity of the designed primers can be assessed using the full distribution of the sequence across the entire *JK* gene plus up- and downstream flanking regions. Moreover, the *JK*A*/*JK*B* critical missense mutation (SNP 838G>A, in exon 9) at chromosomal location ch18:43319519 was found to be heterozygous (50% of each G and A were represented) at a coverage depth of 838X at this location. This SNP encodes for the amino acid substitution Asp280Asn. The genotyping of this SNP in exon 9 matched the serological phenotype provided with the sample (*JK*A*/*JK*B* and Jk a+b+).

A)

B)

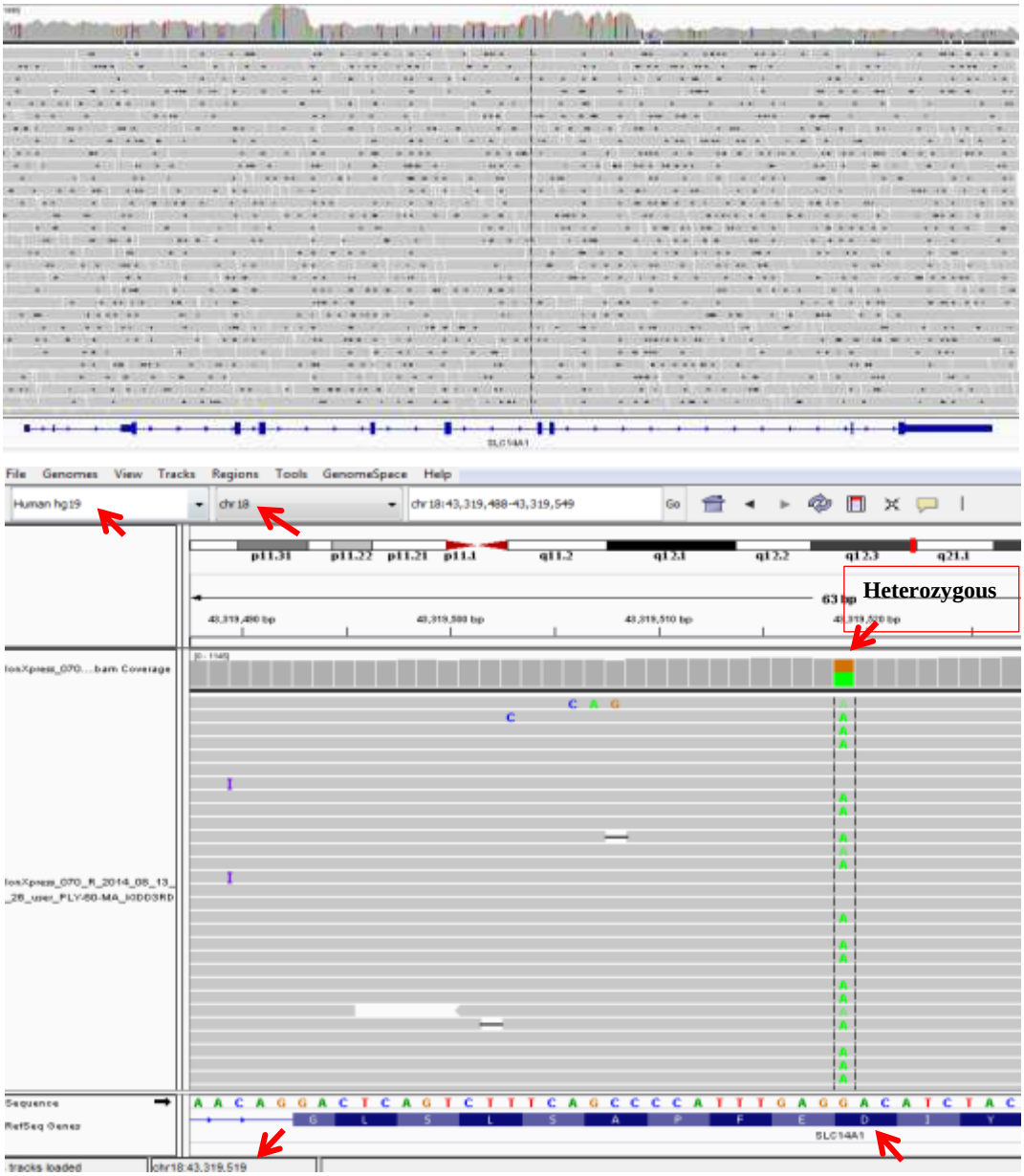


Figure 4.9 An IGV output image showing visualisation analysis of the sequencing data of a Jk (a+b+) sample of *JK*A*/*JK*B* genotype.

A) The reference *JK* gene (*SLC14A1*) (blue) is aligned with the sequencing reads, while the coverage depth is shown above. The coloured bars above represent SNPs across the gene (in exons and introns). Solid bars denote homozygous SNPs, while heterozygous SNPs are denoted by split bars.

B) Red arrows show the critical *JK*A*/*JK*B* missense mutation (SNP 838G>A, in exon 9) at chromosomal location ch18:43319519, which encodes for the amino acid substitution Asp280Asn (GAC>AAC). This SNP is heterozygous, denoted by the split bar (orange/green) for 50 % G/A. Thus, there is a match between the genotype and serology (*JK*A*/*JK*B* and Jk a+b+).

4.3.3.4 Variant analysis and genotyping

The Variant caller plugin of the Ion Suite™ software was used here to analyse the aligned sequence data generated from Ion Suite™ against the reference gene in terms of mutations and zygosity, and was linked to the IGV software (for direct visualising of the mutations). In addition, VCF files obtained from the Variant Caller plugin were used together with other software, such as IGV and SeattleSeq Annotation 137, for confirmation. The VCF files were utilised and uploaded online at SeattleSeq Annotation 137 (SeattleSeqAnnotation137, 2016), which provides a wide range of information, based on the NCBI database, such as the chromosomal location, type of mutation, the transcript number, amino acid change and nucleotide location on the cDNA. Figure 4.10 illustrates an example of the SeattleSeq Annotation 137 output.

inDBSNPOrNot	chromosome	position	referenceBase	sampleGenotype	sampleAlleles	allelesDBSNP	accession	functionGVS	functionDBSNP	rsID	aminoAcids	proteinPosition	cDNAPosition
dbSNP_86	18	43319519	G	R	A/G	A/G	NM_001146036.2	missense	missense	1058396	ASP,ASN	280/390	838

Figure 4.10 An example of variant annotation report on a Jk (a+b+) sample (*JK*A*/*JK*B* genotype) obtained from the SeattleSeq Annotation 137 online tool.

In this example, the *JK*A*/*JK*B* mutation is shown at chromosomal location ch18: 43319519; the reference base is G; the SNP is heterozygous (A/G at 838). The reference transcript number is NM_001146036.2 and the missense mutation encodes Asp280Asn. This SNP is known and has been reported in the dbSNP database (rsID number indicated).

4.3.4 *JK* genotyping by NGS

Extensive genotyping of the *JK* gene was carried out using to NGS, which enabled detection of existing mutations (known and novel). A variety of mutations, including exonic and intronic SNPs, were analysed in 67 *JK* samples. The genotypes were then correlated with the phenotypes.

4.3.4.1 NGS analysis and Genotyping of the *JK* gene

The *JK* gene was fully sequenced in 67 different samples, most of which (59/67) were of known JK serological phenotype. The NGS genotyping data matched with the main predicted phenotypes of *JK* alleles *JK*A*, *JK*B* and *JK*01W*. The exonic SNPs are listed in Table 4.4; a concordance between the crucial *JK*A/JK*B* allele-defining SNP (838G>A) in exon 9 and the Jk^a/Jk^b phenotype (and zygosity) can be seen in all samples with known phenotypes (59 samples). In addition, 10/67 samples were found to carry the SNP 130G>A (encoding for the amino acid substitution Glu44Lys). This mutation has been assigned to *JK*01W* allele (*JK*01W.01*) and is known to weaken Jk^a expression (Wester et al., 2011), although these samples showed normal expression of both Jk antigens (Jk^a and Jk^b). Out of the 67 samples, 9 were heterozygous for the 130G>A SNP and displayed normal expression of the Jk^a antigen: 6/7 had known phenotype Jk (a+b-) and 1/2 had known phenotype Jk(a+b+). One sample with the Jk (a+b-) phenotype was homozygous for this SNP.

The synonymous amino acid substitution Pro196Pro, encoded by the SNP 588A>G in exon 7, was also found in the samples. Genotyping was carried out in four separate runs: in the first two runs, this SNP (588G) was only found in *JK*B* and *JK*01W* (*JK*01W.01*) alleles, while the rest of the *JK*A* samples carried the nucleotide (588A), which suggested a cumulative effect of both 588G and 130A on weakening Jk^a expression. To ensure the *cis* existence of the 588G SNP in both *JK*B* and *JK*01W*

alleles, further investigation was conducted involving sequencing of cDNA clones (section 4.3.5.2). Following further NGS runs, the 588A>G SNP was also found in the *JK*A* allele: 1/18 heterozygous in *JK*A/JK*A*; 1/13 homozygous in *JK*A/JK*B;* and 1/7 homozygous in *JK*A/JK*01W* (Table 4.4). This may suggest that the 588A>G SNP is not involved in the weakening of the Jk^a antigen.

Of the sequenced cohort, 10/67 samples were heterozygous for the 810G>A SNP, which encodes the synonymous amino acid substitution Ala270Ala. This SNP appears to be only carried by the *JK*B* allele, as found in 7/29 *JK*B/JK*B* and 3/13 *JK*A/JK*B* samples. The SNP is located at the exon 8/intron 8 boundary (near the splice region), which may suggest an effect on Jk^b expression. In fact, it was previously suggested that this SNP may abolish the expression of Jk^b antigens and thus represent a novel *JK*B Null* allele (Henny et al., 2014). This SNP is currently not included in the BGMUT database, although it has been reported in the NCBI dbSNP database (dbRBC, 2016, NCBI, 2016a). After cDNA sequence analysis here, no effect of this SNP on Jk^b expression was observed (section 4.3.5.1). A G>A SNP in exon 1, at chromosomal location ch18: 43304182, was noted in all samples carrying the *JK*01W* allele; however, exon 1 is not translated and the SNP results in a synonymous substitution (AGG>AGA), where both nucleotides encode Arg. Thus, this SNP may not affect expression, but rather is specific to the *JK*01W* allele (Table 4.5).

Most of the SNPs observed here required no further confirmation or validation due to high coverage (750X) and previous validation in the NCBI database (dbSNP). However, SNPs 810G>A and 588A>G required further analysis to confirm their associations (sections 4.3.5.1 and 4.3.5.2). With regard to SNPs in the *JK* introns, a high number of SNPs were distributed among the *JK* alleles. On average, there were 30 intronic SNPs in *JK*A/JK*A*, 50 in *JK*A/JK*01W*, 40 in *JK*01W /JK*01W*, 80 in *JK*B/JK*01W*, 70 in *JK*B/JK*B* and 80 in *JK*A/JK*B* samples (data not shown).

NGS defined Genotype (Alleles)	Number of samples sequenced	Sequence Variation				Serological phenotype (No. of samples with phenotype)
		(Exon 9) *JK*A/JK*B* 838G>A Asp280Asn	**(Exon 4) *JK*Aw* 130G>A Glu44Lys**	**(Exon 7) 588A>G Pro196Pro (synonymous substitution)**	**(Exon 8 near splice region) 810G>A Ala270Ala (synonymous substitution)**	
*JK*A/ JK*A*	18	G/G (18)	G/G (18)	A/A (17) A/G (1)	G/G (18)	Jk(a+b-) (17)
*JK*B/JK*B*	26	A/A (26)	G/G (26)	G/G (26)	G/G (19) G/A (7)	Jk(a-b+) (25)
*JK*A/ JK*B*	13	G/A (13)	G/G (13)	A/G (12) G/G (1)	G/G (10) G/A (3)	Jk(a+b+) (9)
*JK*01W.01/ JK*B*	2	G/A (2)	G/A (2)	G/G (2)	G/G (2)	Jk (a+b+) (1)
*JK*01W.01/ JK*A*	7	G/G (7)	G/A (7)	A/G (6) G/G (1)	G/G (7)	Jk(a+b-)(6)
*JK*01W.01/ JK*01W.01*	1	G/G (1)	A/A (1)	G/G (1)	G/G (1)	Jk(a+b-)(1)

Table 4.4. NGS genotyping of 67 samples of differing JK phenotypes.

67 different gDNA samples (59 of which were phenotyped by serology) were sequenced using NGS and *JK*-specific LR-PCR. All phenotyped samples showed complete concordance with the NGS genotyping data with the key *JK*A/JK*B* allele SNP (838G>A). Other exonic SNPs are: 130G>A, suggested to encode for the *JK*01W* allele (*JK*01W.01*); 588A>G, encoding for a synonymous substitution; and 810G>A, claimed to encode for a purported novel *JK*B* null allele. The SNP 810G>A did not appear to affect Jk^b antigenicity as it was expressed in 7 Jk a-b+ and 3 Jka+b+ samples. *Asterisks denote allele names that are the same as in the ISBT (Reid et al., 2012).

4.3.4.2 Assignment of *JK* allele-specific polymorphism patterns

NGS extensive genotyping enabled analysis of both exonic and intronic SNPs, with the latter mainly correlating with *JK*A, JK*B* and *JK*01W* alleles. This allowed detection of allele-specific mutations, including SNPs and a deletion, which may form allele-defining SNP patterns ('fingerprints') thus giving suggested reference sequences for the *JK*A, JK*B* and *JK*01W* alleles. This approach to define allele-specific SNP patterns (reference sequences) involved the comparison of NGS sequences to reference sequences in the human genome (hg19). Sequences of all homozygous-allele samples (for example *JK*A/JK*A* and *JK*B/JK*B*) were then compared to determine their reference sequence. Table 4.5 illustrates the allele-specific reference SNPs (allele-specific) that define *JK*A, JK*B* and *JK*01W* alleles; these show a high degree of concordance to the *JK* genotype at the allele-defining SNP (838G>A, at location ch18: 43319519, in exon 9) that encodes for Asp280Asn. As seen in Table 4.5, there are number of allele-specific SNPs in introns; in the case of *JK*B*, a single nucleotide deletion (G>Del) at location ch18: 43321558. These novel allele-defining patterns have not been described before in the literature. The suggested reference sequence patterns for *JK*A* and *JK*B* alleles are shown here in Table 4.5; multiple *JK*A, JK*B* alleles were immediately recognised and some had to be resequenced to eliminate errors (Table 4.5). Errors were suspected due to their diverse zygosity of SNPs (compared to the proposed reference sequence) and also one sample did not match the phenotype at first (data not shown), which illustrates the feasibility of this approach.

Although the reference sequence from the human genome (hg19) encodes Asp at the 280 position of the *JK*A* allele, multiple intronic SNPs were seen here to be consistent with either *JK*A, JK*B* and *JK*01W*. This implies that the reference sequence from the human genome (hg19) was obtained from individuals of different *JK* genotypes. In addition, most samples (apart from those noted in Figure 4.11 legend) were found to

have a homozygous difference to several nucleotides in the reference sequence form the (hg19); this suggested that these uncommon SNPs (Figure 4.11A) were found during the assembly of human genome sequences. One of these SNPs was 43319359 C>T, located a short distance (160 bp) upstream of the *JK*A*/*JK*B* crucial SNP (838G>A). This SNP is noteworthy to point out, to avoid possible allelic dropout if this position was used within designed genotyping primer in the case of, for example, the current BGG platforms.

All ten (7 *JK*A/JK*01W*, 2 *JK*B/JK*01W* and 1 *JK*01W /JK*01W*) samples shared the *JK*01W* allele as they carried the previously described SNP (130G>A); the 10 samples showed an almost identical pattern (a 'specific fingerprint') (Table 4.5), due to the zygosity of the SNPs defined by the NGS. Interestingly, this approach of extensively analysing allele-specific patterns revealed that the *JK*01W* sequence, in addition to its specific SNPs, appeared to resemble a hybrid sequence of *JK*A/JK*B* alleles. The resemblance of the *JK*B* allele can be seen in part of intron 7, exon 7, parts of intron 9 and intron 10, while the remainder of the *JK*01W* allele appeared to be derived from *JK*A* (see Table 4.5 and Figure 4.11D). Furthermore, an intronic SNP located at position -46 from the 3' end of intron 9 was described to be associated with *JK*A* (nucleotide A) while the *JK*B* and *JK*01W* carry (nucleotide G) (Wester et al., 2011, Irshaid et al., 2000, Daniels, 2013). However, this SNP was found here to be not completely allele-specific as it was heterozygous in 4 *JK*A/JK*A* samples and 1 *JK*B/JK*B* sample whereas G/G in 1 *JK*A/JK*A* sample, 1 *JK*A/JK*01W* sample and 1 *JK*A/JK*B* sample (data not shown in table 4.5). Figure 4.11 provides a graphical representation of these allele-specific SNPs along the *JK* alleles.

Table 4.5 NGS of 67 different *JK* samples, 59 of which were of known Jk phenotype. *(NOTE: the table is on the next page)*.

The table graphically represents the suggested reference SNPs of *JK*A, JK*B and JK*01W (JK*01W.01).* The *JK*A*-specific SNPs (forming a reference sequence and pattern) are shown in red; the reference *JK*B* SNPs and the deletion are shown in blue; and the SNPs defining the *JK*01W* allele (*JK*01W.01*) are shown in green (and are also highlighted in yellow). Homozygous mutations are represented by solid colour, whilst crossed colour represents heterozygous mutations. The major *JK*A/JK*B* SNP (838G>A) is highlighted in orange, while other key mutations (SNPs 588A>G and 810G>A) and the G deletion, at chromosomal location chr: 43321558, are highlighted in grey. Samples with no provided phenotype provided are ND (not defined). The position of each SNP is represented by their location on chromosome 18/hg19. The 3 LR-PCR overlapping amplicons and their location are indicated as black boxes at the top. *Asterisks represent repeat sequence samples for confirmation. There are two apparent *JK*A* alleles (samples JK009. 04, 05, 11, 13 and 15 for one allele while sample JK009.10 for the other) that differ from the reference sequence. Sample JK009.32, sample JK009.39, samples JK009.35, 44 and 53 and sample JK009.52, on the other hand, differ from the reference *JK*B* allele sequence in terms of the intronic polymorphisms. The *JK*Aw* and *JK*01W* alleles are short form for the official ISBT allelic name of *JK*01W.01*. The number 009 was used in accordance with the number of the *JK* blood group system (ISBT number). Larger table in appendix B

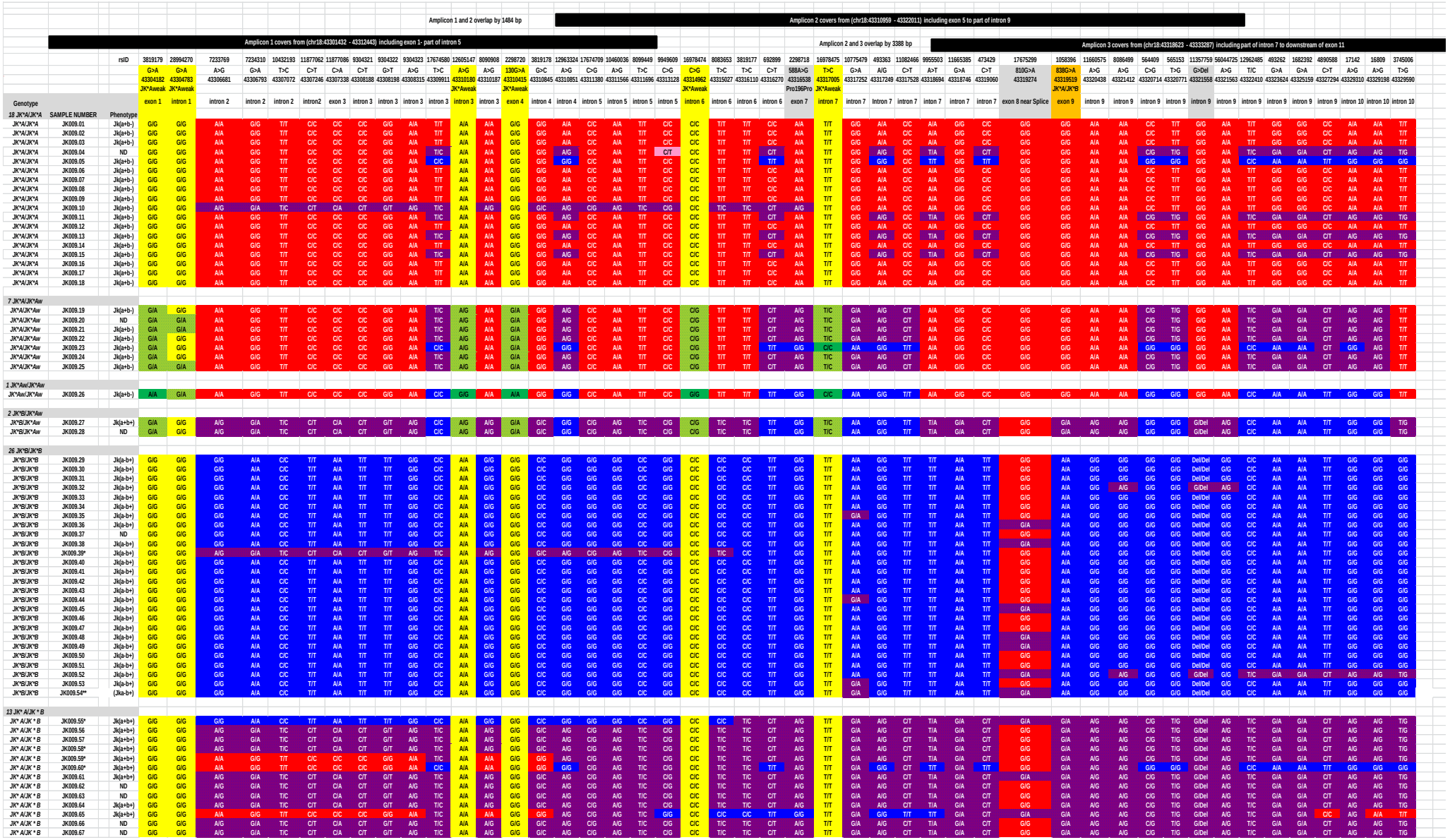
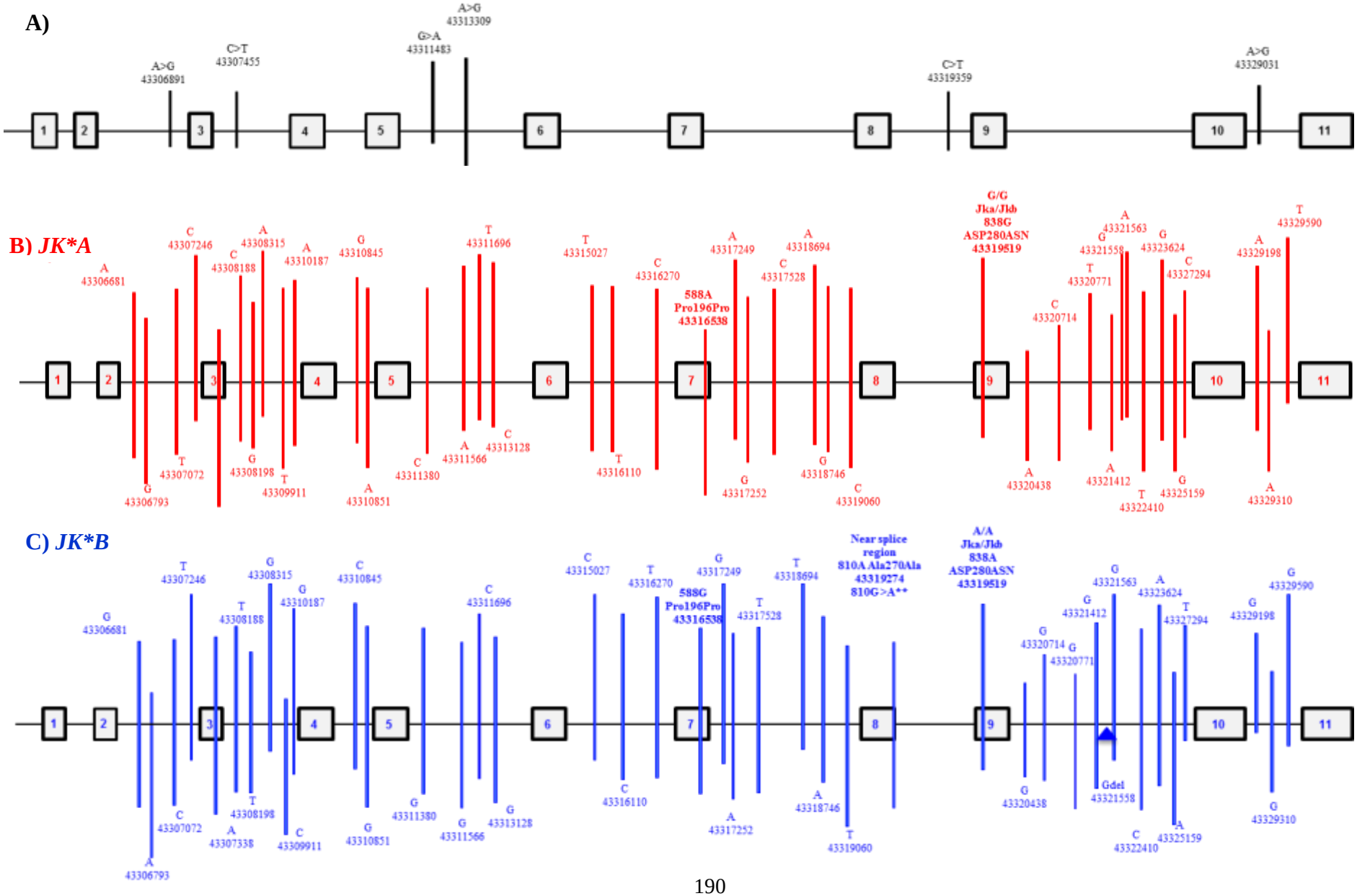
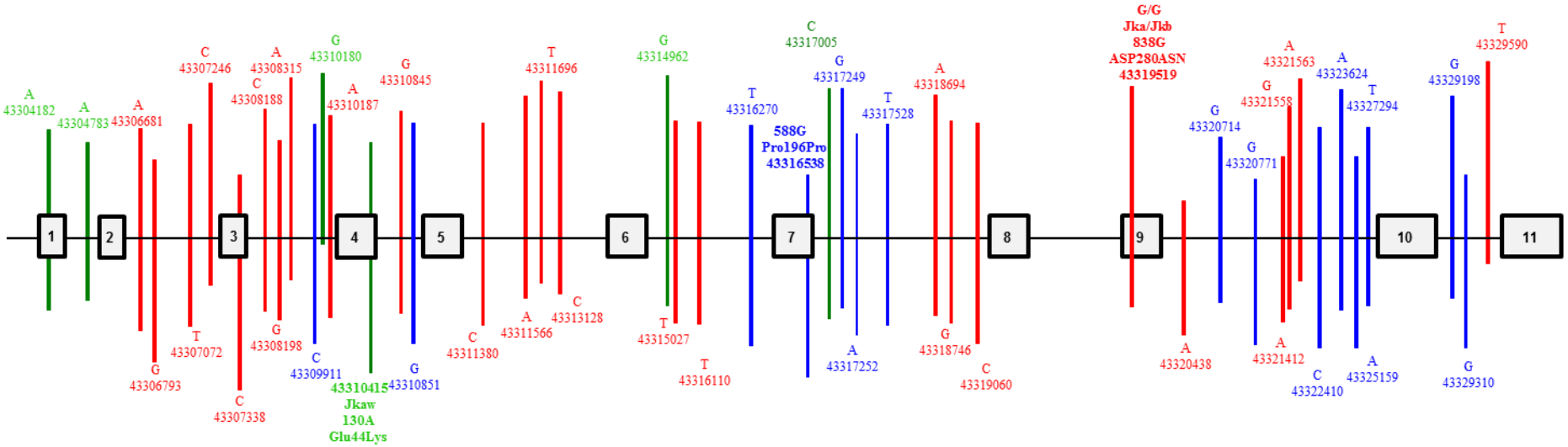

Amplicon 1 and 2 overlap by 1484 bp
Amplicon 2 covers from (chr18:43310959 - 43322011) including exon 5 to part of intron 9
Amplicon 1 covers from (chr18:43301432 - 43312443) including exon 1- part of intron 5
Amplicon 2 and 3 overlap by 3388 bp
Amplicon 3 covers from (chr18:43318623 - 43333287) including part of intron 7 to downstream of exon 11
rsID
Genotype
18 JK*A/JK*A
SAMPLE NUMBER
Phenotype
7 JK*A/JK*Aw
1 JK*Aw/JK*Aw
2 JK*B/JK*Aw
26 JK*B/JK*B
13 JK* A/JK * B

Figure 4.11 Graphic illustration of the allele-specific SNPs forming reference *JK*A*, *JK*B* and *JK*Aw* allele sequences ('fingerprints'). ***(NOTE: the figures 4.11 A, B, C and D are on the next pages)*.**

The genomic structure for the *JK* alleles taken from analysing the 67 different samples derived by NGS is shown here. The SNPs are denoted by vertical lines and their chromosomal locations on ch18: hg19 are shown. **A** shows the SNPs (homozygous) that are different from the reference sequence (hg19) in all samples, apart from those noted here: for example (A>G 43306891), in all samples except the following: (A>G 43306891 (samples JK009.04/11/13/15/60 and 65 were heterozygous, while JK009.05 was homozygous A/A; C>T 43307455 (samples JK009.04/11/13/15/60 and 65 were heterozygous, while JK009.05 was homozygous C/C; G>A 43311483 (samples JK009.04/11/13/15/23/60 and 65 were heterozygous, while JK009.05 was homozygous G/G; A>G 43313309 (samples JK009.04/11/13/15 and 60 were heterozygous, while JK009.05 was homozygous A/A; C>T 43319359 (samples JK009.04/11/13/15/60 and 65 were heterozygous, while JK009.05 was homozygous C/C; A>G 43329031 (sample JK009.65 was heterozygous). **B and C** show the *JK*A* and *JK*B* allele-specific SNPs obtained from analysing the NGS data of homozygous *JK*A* and *JK*B* samples. The majority of samples matched these 'allele fingerprints', which suggests that these sequences may be *JK*A* and *JK*B* reference sequences. The *JK*A* reference SNPs are shown in red while those for the *JK*B* allele are shown as blue lines. Blue arrow denotes the G deletion at chromosomal location 43321558. **D** shows the sequence of samples carrying the *JK*01W* allele, which suggests derivation of SNPs from *JK*A* (red) and *JK*B* (blue). Green lines represent *JK*01W* -specific SNPs. *JK*Aw* and *JK*Aweak* are written as short form for the official ISBT allelic name (*JK*01W.01*). ** The SNP 810G>A was found in ten samples carrying the *JK*B* allele.


A)
A>G
43306891
C>T
43307455
A>G
43313309
G>A
43311483
C>T
43319359
A>G
43329031
1
2
3
4
5
6
7
8
9
10
11
B) JK*A
A
43306681
C
43307246
C
43308188
A
43308315
A
43310187
G
43310845
T
43311696
T
43315027
C
43316270
588A
Pro196Pro
43316538
A
43317249
C
43317528
A
43318694
G/G
Jka/Jkb
838G
ASP280ASN
43319519
A
43321563
G
43321558
G
43323624
T
43320771
C
43320714
C
43327294
T
43329590
A
43329198
G
43306793
T
43307072
G
43308198
T
43309911
A
43310851
C
43311380
A
43311566
C
43313128
T
43316110
G
43317252
G
43318746
C
43319060
A
43320438
A
43321412
T
43322410
G
43325159
A
43329310
C) JK*B
G
43306681
T
43307246
G
43308315
T
43308188
G
43310187
C
43310845
C
43311696
C
43315027
T
43316270
588G
Pro196Pro
43316538
G
43317249
T
43317528
T
43318694
Near splice
region
810A Ala270Ala
43319274
810G>A**
A/A
Jka/Jkb
838A
ASP280ASN
43319519
G
43321563
G
43321412
A
43323624
T
43327294
G
43329198
G
43329590
G
43320714
G
43320771
A
43306793
C
43307072
A
43307338
T
43308198
C
43309911
G
43310851
G
43311380
G
43311566
G
43313128
C
43316110
A
43317252
A
43318746
T
43319060
G
43320438
GdeI
43321558
C
43322410
A
43325159
G
43329310

D) *JK*Aw* (*JK*01W*)


A 43304182
A 43304783
A 43306681
G 43306793
C 43307246
T 43307072
C 43307338
A 43308315
C 43308188
G 43308198
C 43309911
G 43310180
A 43310187
43310415 Jkaw 130A Glu44Lys
G 43310845
G 43310851
C 43311380
T 43311696
A 43311566
C 43313128
G 43314962
T 43315027
T 43316110
T 43316270
588G Pro196Pro 43316538
C 43317005
G 43317249
A 43317252
T 43317528
A 43318694
G 43318746
C 43319060
G/G Jka/Jkb 838G ASP280ASN 43319519
A 43320438
G 43320714
G 43320771
A 43321412
A 43321563
G 43321558
C 43322410
A 43323624
T 43327294
A 43325159
G 43329198
G 43329310
T 43329590
1
2
3
4
5
6
7
8
9
10
11

4.3.5 cDNA analysis of *JK* 810G>A and 588A>G SNPs

4.3.5.1 SNP 810G>A

The SNP 810G>A is thought to encode for the synonymous substitution Ala270Ala (silent mutation), and is located at the second last nucleotide position of exon 8 (at the exon 8/intron 8 boundary) (Figure 4.12). Due to its location, this SNP was previously suggested to alter the expression of the Jk^b antigen by possibly disturbing the splice site (Henny et al., 2014). The authors found this SNP in two samples of Jk b- phenotype, but which carried the *JK*B* allele; this SNP was, therefore, characterised as comprising a novel *JK*B* allele, not previously reported in the BGMUT database (dbRBC, 2016). In contrast to this finding, 10/67 of samples which showed normal Jk^b expression with phenotype (7 Jk a-b+ and 3 Jka+b+) here were found to carry this particular SNP; moreover, these samples showed compatible genotypes (7 *JK*B/JK*B* and 3 *JK*A/JK*B*). Thus, the findings here do not corroborate the previous study by Henny's group (2014), thereby raising the question whether this SNP is indeed a silencing mutation. This was further investigated by examining RNA levels in these samples. For instance, analysis of mRNA would show whether any exon skipping had occurred, which may have led to downregulated or abolished expression of the Jk^b antigen. During these experiments, cDNA was synthesised from the RNA (section 2.3.4 and 2.3.5) of 15/67 samples. The samples were of known phenotype (6 Jk a-b+, 5 Jk a+b+, 2 Jk a+b-, 1 *JK*A/JK*01W* (Jk a+b-) and 1 *JK*B/JK*01W* (Jk a+b+)). Primers were used to amplify the selected areas of interest (exons 8 and 9) to analyse the SNP 810G>A in exon 8 and the (*JK*A/JK*B*) SNP 838G>A in exon 9 (section 2.3.3). This primer design allowed annealing within exons, as introns are naturally spliced out from RNA (Figure 4.13). Figure 4.14 shows the amplified cDNA bands of the expected size (~237 bp). The cDNA amplicons were then sequenced using Sanger sequencing (see section 2.3.6). Only 1/15 samples with the phenotype Jk a+b+, *JK*A/JK*B* was found to be

heterozygous for both the 810G>A and 838G>A SNPs. This sample confirmed there was a restored a splice site sequence, spliced as expected, with no sign of exon skipping that could be silencing the Jk^b antigen, which was manifested by the phenotype Jk b+ (Figure 4.15). The data generated by Sanger sequencing matched with NGS data for *JK*A/JK*B* and serology (Jka+b+). Notably, the 5' exonic reference sequence of *JK* at the exon/intron boundary appeared to be a GG-intron, which might not match the consensus splice site sequence (AG-intron), as shown in Figure 4.16.

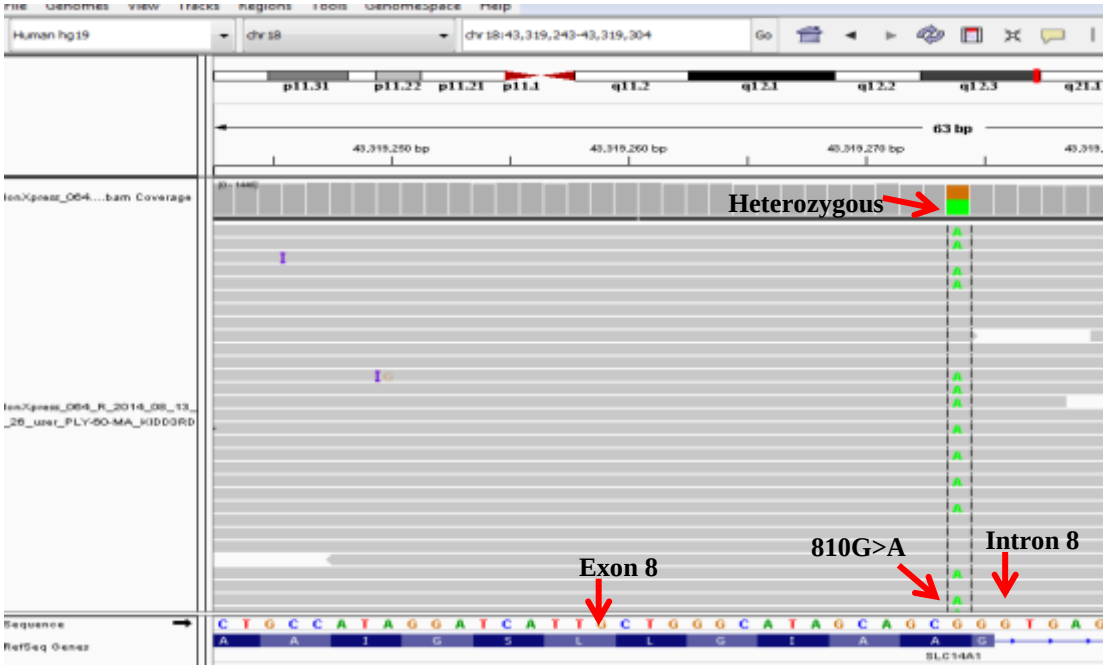


Figure 4.12 An IGV image illustrating the critical location of the 810G>A SNP in a *JK* sample.

The SNP 810G>A is located at the exon 8/intron 8 boundary; the 3' of exon 8 within the 5' splice site sequence. The location of the SNP is denoted with a split box (orange and green) due to its heterozygosity (G>A).

Figure 4.13 Primer design for amplification of SNPs 810G>A and 838G>A.

The above primer pairs were used: forward (red brackets) and reverse (blue brackets) to address the 810G>A and 838G>A SNPs. These primers were designed to anneal within exons instead of introns, as the latter are spliced out in mRNA synthesis. The forward primer annealed within exon 8 to cover 810G>A (red circle), while the reverse primer annealed within exon 9 to cover 838G>A (black circle).

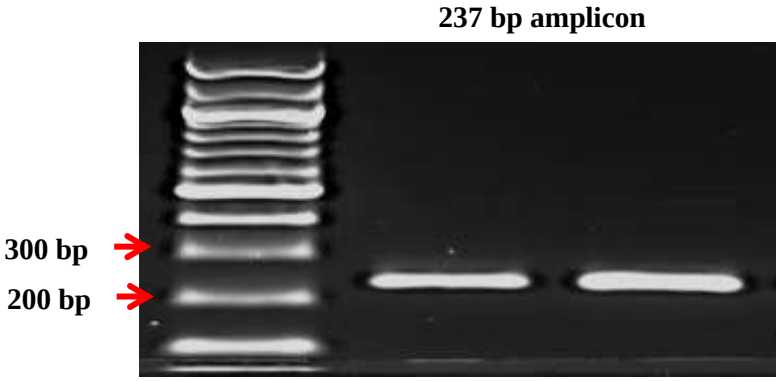


Figure 4.14 The amplification of *JK* cDNA for splice site analysis

Amplification of *JK* cDNA from two samples (JK009.55 and 34, respectively) is shown here as an example under the thermocycling conditions described in section 2.3.5. The amplicon size is 237bp, seen just above the 200bp ladder. Samples were loaded onto a 1.5% agarose gel and electrophoresed at 90V for 1 hour. The TriDye™ 100bp was used as a marker of amplicon size.

Sample JK009.55 (Jk a+b+, *JK*A/JK*B*)

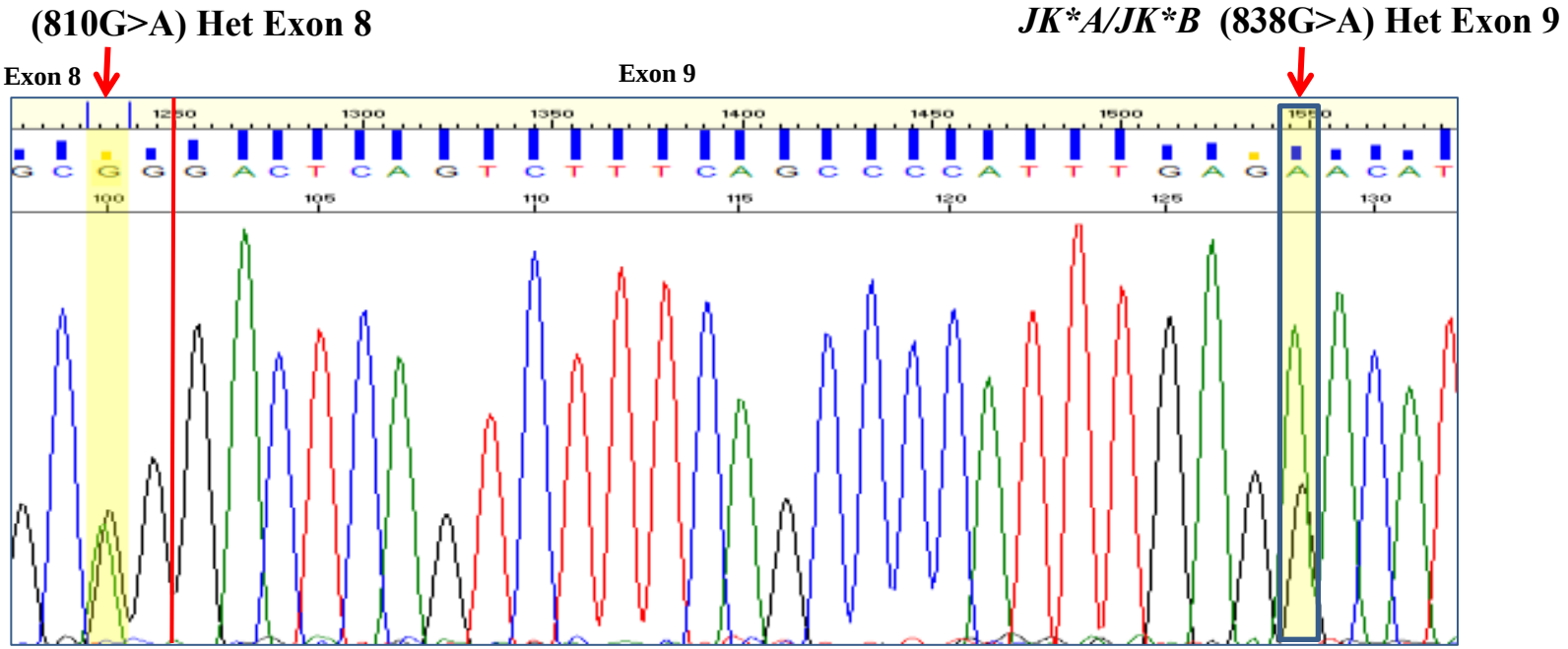


Figure 4.15 An electropherogram of 810G>A and 838G>A SNPs from Sanger sequencing of *JK*A/JK*B* cDNA.

The sample JK009.55 was heterozygous for both the 810G>A and 838G>A SNPs. The introns between both exons 8 and 9 were visibly normally spliced out, with no sign of exon skipping; thus, both SNPs were expressed and validate the provided phenotype (Jk a+b+) and genotype by NGS (*JK*A/JK*B*), without affecting the Jk^b expression. Image was obtained by Seq scanner 2. A red line split between exons 8 and 9.

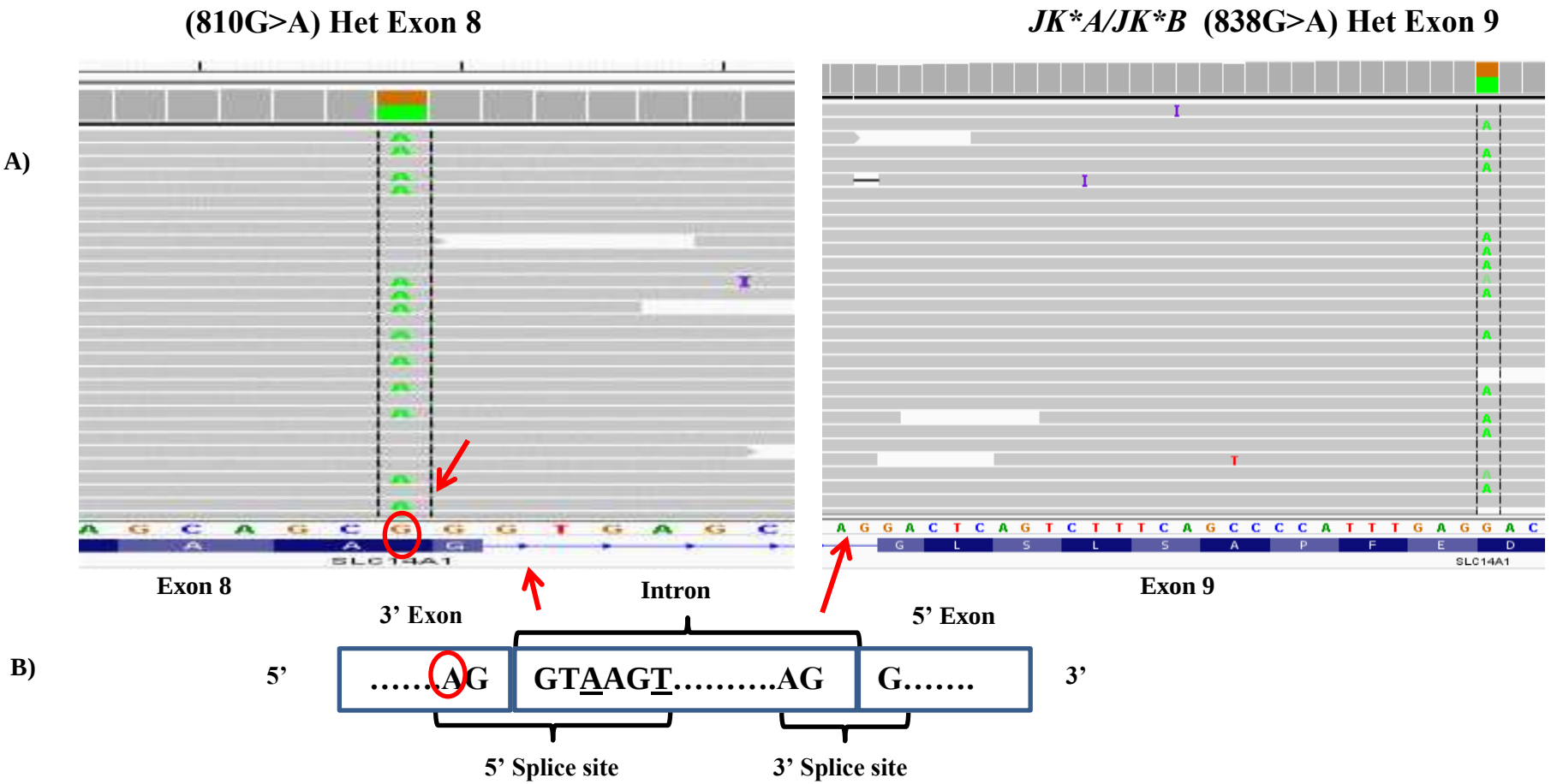


Figure 4.16. An IGV image of the 810 G>A and 838 G>A SNPs and shows the splice site sequence between exon 8 and 9 of a *JK*A/JK*B* sample.

A) The 3' exonic reference sequence of *JK* at the exon 8/intron 8 boundary appeared to be GG-intron, which may not match the consensus sequence (AG-intron) shown in B) red circle. B) the consensus splice site sequence is shown, where the underlined A can change to G. In addition, the consensus T in the intron is found as C in the *JK* reference sequence. Both SNPs are heterozygous in this sample. Encoding for the genotype *JK*A/JK*B.*

4.3.5.2 SNP 588A>G

Sanger sequencing was used to sequence cDNA clones from *JK* samples carrying *JK*B/JK*01W* alleles (JK009.27 Jka+b+), using RT-PCR and primers annealing within exons 4-9 (Figure 4.17 and 4.18). This was conducted to ascertain whether the G588 silent SNP was carried by both *JK*B* and *JK*01W* alleles as suggested by NGS genotyping data (Table 4.4 and 4.5), but not with the *JK*A* allele (the first two runs). As the NGS approach uses small fragments (200-400bp), assigning mutations as *cis* or *trans* (i.e. depending on which allele a mutation arises from compared to other mutations) is more difficult than with cloning methods, where plasmid DNA take only one cDNA molecule (in this case, cDNA of only one allele). The successful insertion of cDNA into plasmid DNA was confirmed using *EcoRI* restriction digest, where both uncut and cut plasmid DNA was seen (Figure 4.19). The Sanger sequence analysis confirmed the findings of NGS; that both alleles (*JK*B* and *JK*01W*) carry G588, while A838 and G130 were carried by the *JK*B* allele, and G838 and A130 were carried by the *JK*01W* allele. The nucleotide G810 in exon 8 was found in both alleles, which also confirms NGS data in that sample (Figure 4.20). This 588A>G SNP might not be involved in suppression of the Jk^a antigen as it was found in multiple samples carrying the *JK*A* allele (1/18 het in *JK*A/JK*A*, 1/13 hom in *JK*A/JK*B* and 1/7 hom *JK*A/JK*01W*) (Table 4.4 and 4.5). However, it may play cumulative role with other polymorphisms.

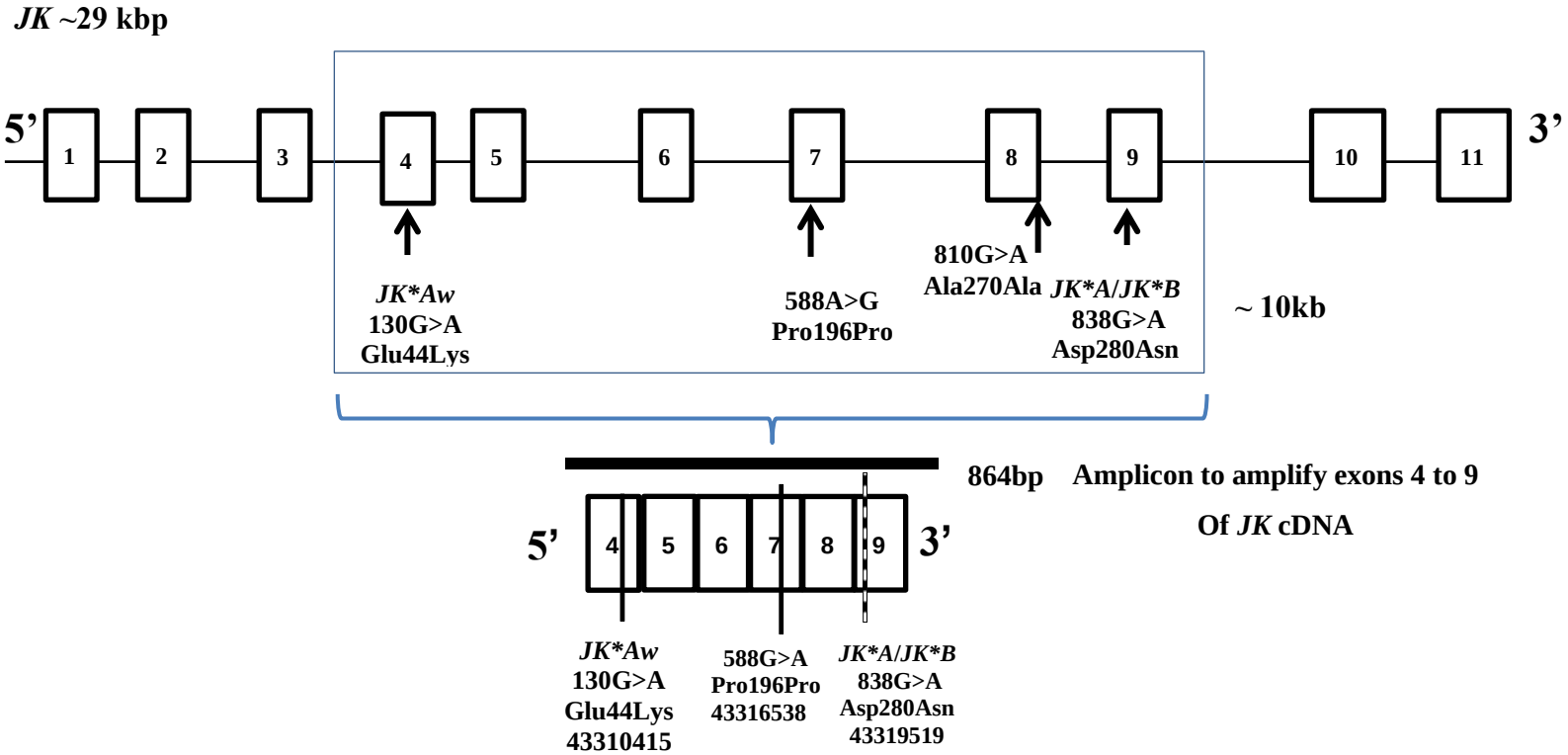


Figure 4.17 The primer pairs designed to cover exon 4 to exon 9 for 588A>G SNP analysis.

The primer pairs were designed to produce a cDNA amplicon of 864bp in size, which covers the following SNPs: 130G>A (*JK*01W* specific); 588A>G; and 838G>A (*JK*A/JK*B* specific). This analysis determined whether 588A>G was associated with both *JK*01W* and *JK*B* alleles. The SNP 810G>A in exon 8 was also covered. The total distance across exons 4 to 9 is ~10kb in the *JK* gene which was covered in the NGS and LR-PCR analysis. *JK*Aw* is another name form for the allelic name (*JK*01W*).

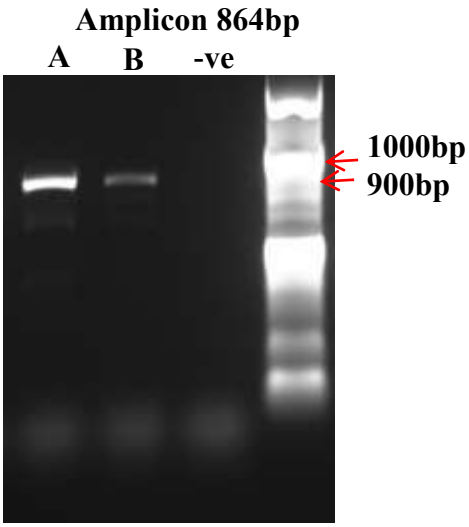
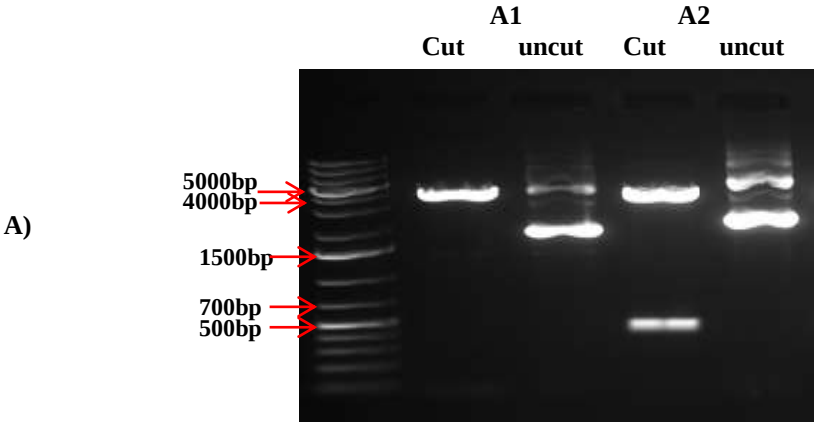


Figure 4.18 *JK* cDNA amplicon (864bp) containing SNP 588A>G.

The amplification of *JK* cDNA from 2 samples is shown here as an example (only sample A was used for downstream analysis due successful colony growth) under the thermocycling conditions (in section 2.3.5). The amplicon size is 864bp, seen around the 900bp ladder. Samples were loaded onto a 1% agarose gel and electrophoresed at 90V for 1 hour. The TriDye™ 100bp was used as a marker of amplicon size.



B)

```
TGGTTAGGGGTGAAAACCAGgtttcgccatgtcaagggagaaggtgcttccccaaagctc
ttggctatgtcaccggtgacatgaaagaacttgccaaccagcttaaagacaaacccgtgg
tgctccagttcattgactggattctccggggcatatcccaagtggtgttcgtcaacaacc
ccgtcagtggaatcctgattctggtaggacttcttgttcagaacccctggtgggctctca
ctggctggctgggaacagtggtctccactctgatggccctcttgctcagccaggacaggt
cattaatagcatctgggctctatggctacaatgccaccctggtgggagtactcatggctg
tcttttcggacaagggagactatttctggtggctgttactccctgtatgtgctatgtcca
tgacttgcccaatttttctcaagtgcattg^aattccatgctcagcaaatgggacctccccg
tcttcaccctccctttcaacatggcgttgtcaatgtacctttcagccacaggacattaca
atccattctttccagccaaactggtcatacctataactacagctccaaatatctcctggt
ctgacctcagtgccctggagttgttgaaatctataccagtgggagttggtcagatctatg
gctgtgataatccatggacaggggcattttcctgggagccatcctactctcctccccac
tcatgtgcctgcatgctgccataggatcattgctgggcatagcagcgggactcagtcttt
cagccccatttgaggacatctactttggactctgggggtttcaacagctctctggcctgca
ttgcAATGGGAGGAATGTTCATGG
```

Figure 4.19 *EcoRI* restriction enzyme analyses for positive insertions.

The cut plasmid DNA (linearised) is aligned along the 4kb marker (~ size of plasmid DNA used), while the uncut fragment migrated faster through the gel due to its circular compact shape (with some remaining open circular plasmid on top of the uncut due to handling). The GeneRuler™1Kb Plus DNA ladder was used here as a marker of DNA size.

A1) A plasmid DNA form a colony with unsuccessful cDNA insertion as no product was shown with *EcoRI* restriction digest. **A2)** cDNA was successfully inserted but the size of the insert was smaller (~500bp) than expected (864bp). **(B)** This was due to the *EcoRI* recognition sequence (G^AATTC) located in the middle of the insert, dividing it into 449bp and 415bp fragments before and after the cut.

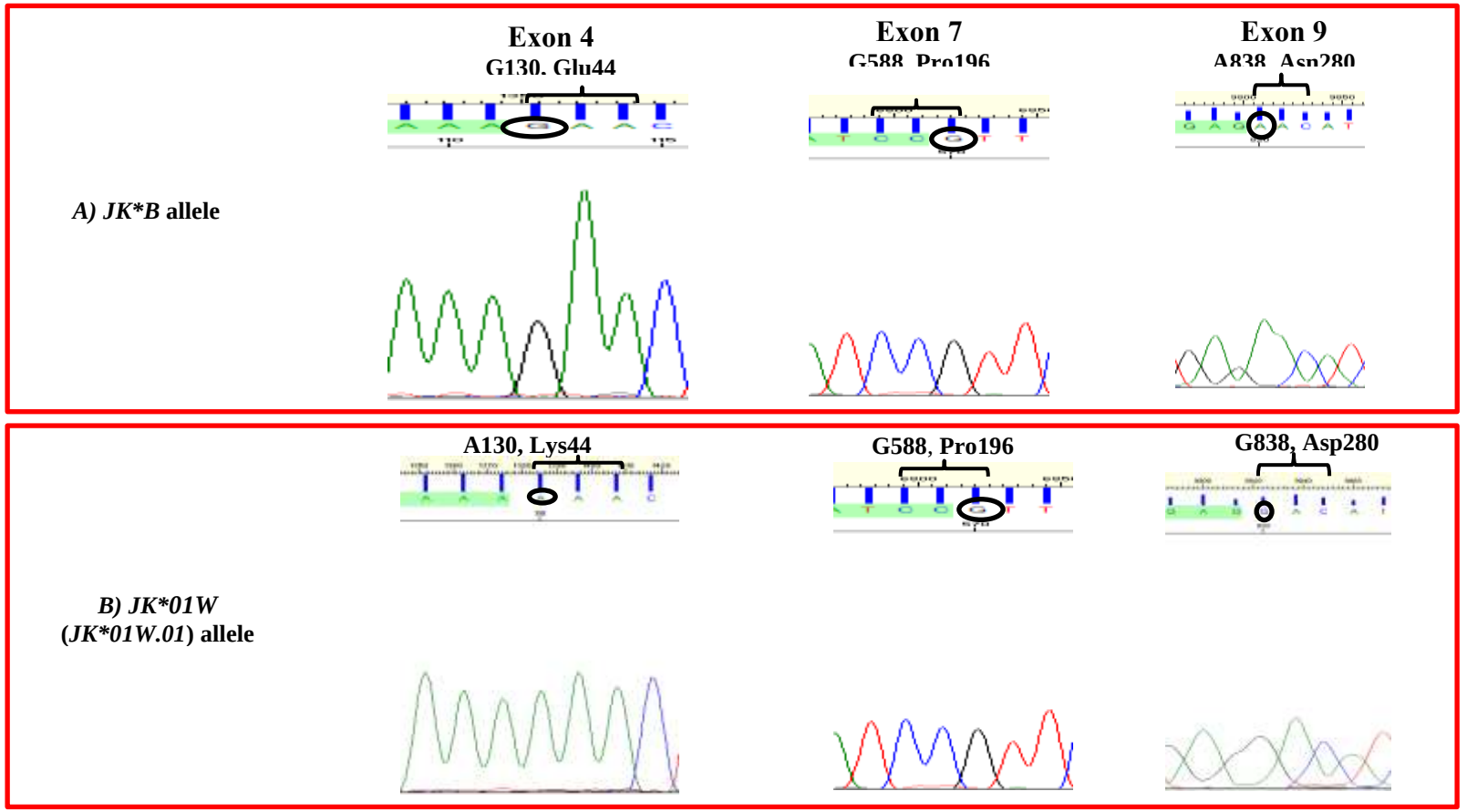


Figure 4.20 Electropherograms showing the association of SNP G588 with both *JK*B* and *JK*01W* alleles by Sanger sequencing of *JK*B/JK*01W* cDNA clones.

A) cDNA clone of the *JK*B* allele confirmed presence of G588 in exon 7, G130 in exon 4 and A838 (specific for *JK*B*) in exon 9. **B)** That from the *JK*01W* allele also confirmed presence of G588 in exon 7, along with A130 (*JK*01W.01*-specific) and G838 (specific for *JK*A*) in exon 9.

4.4 Discussion

4.4.1 NGS of the *JK* library

The construction of *JK* libraries for the NGS genotyping protocol involved various optimisation steps, in particular the LR-PCR conditions, purification and size-selection. The SPRIselect® reagent kit (Beckman Coulter, UK) was preferred for purification and size-selection over AMPure® XP beads Reagent and the Pippin Prep™ instrument due to its cost-effectivity and time efficiency (it can be used for both purification and size selection) (see section 2.2.4.10). Briefly, each Pippin Prep™ kit, which contains ten cassettes, is only capable of processing 4 samples per cassette and run, which takes about 1 hour and ten minutes. The kit costs £461 for size selecting 40 samples (£11.53 per sample) that in addition needs purification steps with AMPure® XP beads Reagent. On the other hand, SPRIselect® reagent kit, which is used for size selection and purification simultaneously, costs £696 for 60 ml, by which 120 samples can be processed (£5.80 per sample).

4.4.2. Quality control of NGS data

Quality control of NGS data was carried out to so as to assess the validity of data for downstream analysis. FastQC was used, which assesses sequence quality by providing a report of the library, based on a Phred score (Andrews, 2016). All *JK* sequences were shown to have high quality (according to the Phred score), with > 99% base call accuracy (1 in 1000 probability of an incorrect base call). In addition to this high quality score, a high mean coverage depth of 750X was shown, which increased the confidence in the data for downstream analysis, such as variant analysis (Sims et al., 2014). This coverage depth significantly exceeded that of previous high-throughput sequencing studies, such as those by Stabentheiner et al. (2011) and Fichou et al. (2014) groups. These studies relied on a 50X coverage depth for polymorphism analysis of the *RHD* gene and 15 blood group system genes, respectively. One explanation for the high

coverage depth seen here could be the use of only a fraction of the 316^{TM} chip capacity in each run; around 108 samples can be simultaneously sequenced for the entire J*K* gene and flanking regions, with total amplicon size (~37kb) in a single run using the 200 base reads with 50X coverage depth as used by Stabentheiner et al. (2011). Moreover, about 540 *JK* samples can be sequenced in parallel per run if the 318^{TM} chip was used (applying the calculation in chapter 1). Using these chips allows simultaneous sequencing of a large number of samples, which drops the cost of sequencing. The cost of the NGS library preparation, which included three PCR reactions (for the 3 different amplicons), fragmentation, purification and quality check by Bioanalyzer 2100, for one *JK* sample was approximately £63.00. On the other hand, the cost of the sequencing run, using the 316^{TM} chips and including template preparation, was around £454.00. Therefore, the cost for sequencing one *JK* sample was approximately £4.20, which would be lower if a higher capacity chip was used.

4.4.3 Validity of *JK* NGS genotyping.

For the majority of the *JK* polymorphisms, no further validation approaches were conducted due to the high quality of data. The base call accuracy was above 99% (~30 Phred score), which was higher than the score of a previous study by Ajay's group (2011; 20 Phred score), which aimed to establish a sequencing guide for data quality. Moreover, the high coverage depth supported the accuracy of the data. Moreover, the important *JK*A/JK*B* SNP (838A>G) observed from NGS showed complete concordance with the provided serology of the Jk$^{a/b}$. Sanger sequencing, conducted to analyse two cases (sections 4.3.5.1 and 4.3.5.2), with SNPs 130G>A, 588A>G, 810G>A and 838G>A involved, also showed complete concordance with NGS data, which confirms its validity.

4.4.4 *JK* NGS-based genotyping

Genotyping of the *JK* gene is suggested to be beneficial for elucidation of the molecular mechanisms of the various *JK* alleles, a number of which weaken or silence the Jk antigens (Reid et al., 2012). This also allows for more accurate prediction of the phenotypes resulting from these alleles, which should help prevent alloimmunisation – especially in transfusion-reliant individuals (Anstee, 2009, Lomas-Francis, 2007). Despite their high throughput capability, the current platforms that are based on microarray technology predict the phenotypes only of predefined SNPs (Hurd and Nelson, 2009) and are not able to detect rare or novel SNPs (not included in the platform) that may affect Jk antigenicity, leading to false interpretation of phenotype. One example of this is the previous study by Paris' group (2014), in which a 96-well flexible microarray platform was designed for automated genotyping of four blood group systems (KEL, JK, FY and MNS). Despite the advantage of automation, which required only 8 hours to process 96 samples, the predicted phenotype of 3/960 samples (involved the KEL, JK and MNS systems) were revealed discordant with serology. The *JK* samples here were genotyped as *JK*A/JK*B* while the phenotype was Jk (a+b-), which was assumed that this was accounted for by a possible silencing polymorphism that disrupt the expression of Jk^b antigen that clearly was not included in the panel list of SNPs. Only 8 SNPs, including *JK*A/JK*B* 838 G>A were included for the 4 blood group systems, with the cost of genotyping of each SNP was below $2.60 (Paris et al., 2014). Another example that displays the limitation of microarray-based high-throughput genotyping, in terms of the restriction to predefined SNPs, was that of the Keller et al. (2014) study. In the study, disagreement between the serological phenotype and that predicted using HEA BeadChip (microarray-based platform) was demonstrated in 7/>60,000 *JK* samples (5 with Jk^b and 2 Jk^a). These samples were further investigated, using DNA sequence analysis, in an attempt to resolve the mismatch. The phenotype of

all 5 samples with the Jk^b discrepancy were predicted from the genotyping to be Jk(b+), while the serology demonstrated no expression of the Jk^b antigen (Jk(b-), which appeared to be due to *JK*B* null alleles, following DNA sequencing analysis. Three samples carried the *JK*02N.01* allele with the polymorphism IVS5-1g>a (one was homozygous), one carried *JK*02N.10* (194G>A) and one carried *JK*02N.09* (191G>A). The two samples showing a Jk^a mismatch were found to carry the *JK*01W.01*-specific SNP 130G>A, and a novel SNP (814>T, Leu272Phe) in exon 9. The latter was thought to hinder the PCR amplification or hybridisation to the *JK*A/JK*B* SNP (838G/A)-specific probe, leading to the allelic drop. Although the number of the samples showing discrepancy was low (7/60,000), it is still considered a limitation of the microarray-based genotyping when screening for rare or new SNPs, while the sequencing-based genotyping was used to resolve this issue that illustrated the advantages of the discovery mode (Keller et al., 2014). Another discrepancy between the serological phenotype of the Jk antigen and (HEA) BeadChip microarray-predicted phenotype was described by Casas and colleagues (2015). This retrospective study evaluated the importance of genotyping over serology in assigning the phenotype of 494 multiply-transfused (SC) patients and limiting alloimmunisation by comparing the historical phenotype with that predicted by microarray screening. One sample was phenotyped as Jk(b-) by both historical and repeat serology; however, the microarray-predicted genotype was Jk(b+), which was subsequently confirmed by sequencing and revealed that the *JK*B* allele carried the silencing 191G>A SNP. This suggested that this SNP was not covered by the microarray panel (Casas et al., 2015). As a result, it can be concluded that sequencing-based genotyping is superior to microarray platforms in terms of providing more extensive genotyping and, thus, better prediction of the true phenotype. This project features the first NGS-based genotyping study of the *JK* gene, coupled with LR-PCR. Unlike other NGS genotyping approaches, such as AmpliSeq^{TM,} in which mostly the

coding regions are targeted, the protocol developed here provides comprehensive genotyping of all existing polymorphisms in the whole *JK* gene (exons, splice sites and introns) plus flanking regions, enabling better assessment of the gene. In addition, the size of the LR-PCR amplicons allows better association between the carried SNPs – particularly novel SNPs – and critical allele SNPs (especially the *JK*A/JK*B* SNP 838G>A); i.e. *cis/trans* association, since LR-PCR amplicons are amenable to cloning (section 4.3.5.2). Moreover, the suggested reference *JK* alleles sequences were established here based on extensive genotyping of both exons and introns by NGS (section 4.3.4.1).

4.4.4.1 Genotyping of *JK* SNPs in exons

Out of the 67 samples, 59 were of known serological phenotype and were in concordance with the NGS genotyping data of key *JK* allele SNPs (*JK*A/JK*B* 838G>A, and *JK*01W* 130G>A). The variant analysis, which included SNP and amino acid locations, was based on the *JK* reference sequence associated with NCBI accession number NM_001146036.2; in contrast to that suggested in Reid et al (2012) with accession number NM_015865. The reason for this is the corroboration between the studies showing that *JK* gene contains 11 exons (Westhoff and Reid, 2004, Lucien et al., 1998) and the 11 exon reference sequence NM_001146036.2; on the other hand, the NM_015865 reference sequence revealed only 10 exons. In addition, the location of the two critical SNPs (*JK*A/JK*B* 838G>A and *JK*01W.01* 130G>A) in the reference sequence used here were in concordance with the locations previously described (exon 9 and 4, respectively) (Daniels, 2013, Wester et al., 2011), unlike in the transcript variant associated with NM_015865 (exon 8 and 3). Despite these differences, the NCBI database confirms that both reference sequences encode the same isoform of Jk antigen

(urea transporter) (NCBI, 2016b).

SNP 130 G>A, described to be *JK*01W.01*-specific and suppress Jk^a expression (Wester et al., 2011), was shared among 11/134 alleles, at a frequency of around 8% - double that previously observed in 300 Swedish samples (Wester et al., 2011). However, the provided serological phenotype showed normal expression of Jk^a, which may be due to the inclusion of weak expression into the positive category. However, in the study by Wester et al. (2011), other factors than this SNP (130G>A) were suggested to reduce expression of the Jk^a antigen, as two samples (homozygous and heterozygous 130G>A) were shown to react strongly in the haemagglutination assay. Nevertheless, this SNP may play a role in Jk^a suppression, as six of samples carrying this SNP showed decreased reactivity in the haemagglutination assay and two, tested with flow cytometry, showed lower expression of Jk^a antigens. However, it is possible that other SNPs located in regions (exonic or promoter) not covered in the previous study may have exerted an effect on Jk^a expression. A silent mutation (588A>G) was found to be carried in all NGS-analysed samples containing *JK*01W.01* (*JK*01W*) alleles (11 haplotypes); this mutation is suggested to also be *JK*B*-specific, along with an intronic SNP at position -46(G) from the 3' end of intron 9 (Wester et al., 2011, Irshaid et al., 2000, Daniels, 2013). Results from Sanger sequencing the cDNA clones confirmed these findings; however, the *JK*01W* was concluded to differ from the *JK*A* allele by at least three SNPs: (130G>A), and two SNPs probably specific to *JK*B*. The SNP 588A>G was found also carried by the *JK*A* allele (3 haplotypes); thus, it appeared to be neither weakening expression of Jk^a antigen nor *JK*B*-specific. Similarly, the SNP in intron 9 (position -46 G from the 3' end) was found heterozygous in 4 *JK*A/JK*A*, G/G in 1 *JK*A/JK*A*, *JK*A/JK*01W* and 1 *JK*A/JK*B* sample. Another SNP in exon 1 (G>A), at chromosomal location ch18: 43304182, was found in all samples carrying the *JK*01W* allele (Table 4.5; however, due to its location in the untranslated exon (1), the

SNP did not encode a ATG codon, but rather, produced a synonymous substitution (Arg-Arg) which did not to affect expression. However, this substitution may have played a cumulative role, along with the other SNPs, in suppressing Jk^a antigen expression.

SNP 810G>A was previously found in the second last nucleotide in exon 8 (at the exon 8/intron8 boundary) and was suggested to exert a silencing effect on the expression of Jk^b antigens (Henny et al., 2014). The authors found this SNP in two (Jk a+b-) samples and defined it as a novel *JK*B Null* allele, since the SNP 810G>A was not already recorded in the BGMUT (Henny et al., 2014). However, NGS sequencing of *JK* samples here revealed that this SNP had no effect on Jk^b antigen expression, as the predicted phenotype (which matched serology) showed positive expression in all 10 samples carrying that SNP. In addition, cDNA Sanger sequencing confirmed this, as no effect of the SNP was found (section 4.3.5.1) and intron 8 was spliced correctly, while splice site sequence seen in exon8/intron 8 boundary was described before (Lucien et al., 1998). In conclusion, this reported novel *JK*B Null* allele (carried in *JK*B*; Henny et al., 2014) is shown here to have no effect on expression of Jk^b antigen.

4.4.4.2 *JK* polymorphism patterns and assignment of *JK*A*, *JK*B* and *JK*01W* allele reference sequences ('fingerprints')

The extensive sequencing of the complete *JK* gene by NGS enabled the analysis of intronic SNPs and their correlation with key allele-defining exonic SNPs, such as *JK*A/JK*B* (838G>A) and *JK*01W* (130G>A). Despite the large number of intronic SNPs found here, those closely correlated with key exonic SNPs in homozygous *JK*A/JK*A, JK*B/JK*B* and *JK*01W/JK*01W* samples were thoroughly analysed. This allowed detection of novel allele-defining SNPs (which included G deletion 43321558 in *JK*B*), based on which suggested allele-specific patterns ('fingerprints') of *JK*A*,

*JK*B* and *JK*01W* reference sequences were established (Table 4.5 and Figure 4.11). The benefit of this assignment of unique patterns provides insight into the evolution of alleles and their variants (intronic or exonic); for example the discovery of novel SNPs from which *JK* alleles may have arisen. This approach of analysing and associating intronic SNPs to alleles yielded interesting findings regarding the *JK*01W* allele, the sequence of which appeared to evolve from or resemble an undescribed hybrid *JK*A/JK*B* gene sequence – a phenomenon that has been described before in other blood group systems, such as ABO (Storry and Olsson, 2004), (Table 4.5 and Figure 4.11). The *JK*01W* allele specific polymorphisms (illustrated in green in Table 4.5 and Figure 4.11) might have evolved on a hybrid *JK*A/JK*B* backbone. In addition, a number of *JK*A* and *JK*B* alleles in samples (Table 4.5) showed some degree of diversity from the suggested reference sequence to generate multiple alleles. This comprehensive analysis showed that the human genome (hg19) may contain uncommon variants, since most of the samples were homozygous different (legend of Figure 4.11). Also, it is worth noting that one of those SNPs (43319359 C>T) is closely located (160bp) upstream of the *JK*A/JK*B* key SNP (838G>A), possibly accounting for the allelic dropout in genotyping platform based on primers designed at this position, such as microarray platforms. Accordingly, this would urge caution when designing *JK*A/JK*B* genotyping primers to avoid amplification failure due to SNPs in the primer binding sites. In addition, it is suggested to be good practice to sequence a large number of samples using NGS for different blood group genes to investigate and catalogue polymorphisms close to critical SNP positions so that these can be taken into consideration whilst designing genotyping primers. One example of allelic dropout due to intronic SNPs was described in a study by Mullins et al. (2007), the goal of which was to develop a sequencing assay for genotyping of cadherin 1 (*CDH1*) gene. The authors reported that the allelic dropout, characterised by a (2398delC) in exon 15, was

because of a SNP located in intron 15 within the primer binding site (Mullins et al., 2007). As a consequence, this would provide a false negative result, which was also reported by another study that reported the possibility of allelic dropout due to intronic SNPs that prevent the discovery of long QT syndrome (LQTS) causing mutations (Tester et al., 2006). Therefore, polymorphisms in introns should be taken into account during designing primers.

In summary, NGS showed clear superiority over other high-throughput genotyping microarray-based methods, since all existing polymorphisms, without the predefined SNPs constraints, were revealed here. SNPs in exons were correlated here with intronic SNPs, which was beneficial in studying the phenotype and evolution of the blood group gene system. In addition, alleles differing form the suggested reference sequences may require further attention for their role in antigenicity. These allele-specific fingerprints may be shared with software developers to establish more detailed interpretation of the genotype and predicted phenotype of *JK* and other blood genes. Moreover, this approach would greatly help solve unusual data and discrepancies in samples, so as to enable correct identification of causative SNPs. Finally, NGS-based genotyping of the *JK* gene has provided insight into the molecular basis of Jk antigen expression, which can be applied to other blood group systems, contributing greatly towards the safety of the blood transfusion. This will be discussed with respect to the ABO group in the next chapter.

Chapter 5

Genotyping of the ABO Blood Group by Next Generation Sequencing

5.1 Introduction

ABO (ABO/ISBT 001) is considered the most clinically significant blood group system in transfusion medicine and transplantation, as naturally occurring ABO antibodies can induce a rapid transfusion reaction if the mismatched ABO blood type is transfused into a patient (Yamamoto, 2004). The *ABO* gene, which comprises 7 exons (Yamamoto et al., 1995), indirectly generates synthesis of A and B antigens by encoding the glycosyltransferases: N-acetylgalactosaminyltransferase (GTA) and galactosyltransferase (GTB), which catalyse expression of these antigens (Daniels, 2013). *ABO* is one of the most complex and polymorphic blood group genes with an ever-increasing number of alleles as in four years from 2012 (Patnaik et al., 2012) to 2016, an increase of 109 new *ABO* alleles have been identified and listed in the BGMUT database (dbRBC, 2016), with many others likely remaining unidentified nor listed (Lang et al., 2016). These alleles have resulted from various polymorphisms – such as, single nucleotide polymorphisms (SNPs) and deletions – within exonic or other parts of the *ABO* gene, including intronic regions. Notably, a number of these polymorphisms might not only alter the specificity of the enzymes but also suppress expression of ABO antigens (Sano et al., 2012, Reid et al., 2012). ABO serological typing has yielded some discrepancies, for which reason genotyping has been used as a more reliable tool for prediction of the ABO phenotype, thereby ensuring safe blood transfusion. Genotyping of the whole *ABO* gene (molecular information across the exons and introns) allows identification and assessment of hybrid, rare and novel alleles (Thuresson et al., 2012, Huh et al., 2011) and allele evolution studies (Avent et al.,

2015). Moreover, a high-throughput genotyping approach like NGS is required to address the increasing number of new *ABO* alleles identified in individuals. The commercially available high-throughput genotyping platforms, such as microarrays, have the disadvantage of only assessing known polymorphisms, thereby preventing discovery of new or rare mutations and resulting alleles (Tilley and Grimsley, 2014). In contrast, NGS is a powerful genotyping platform which enables high-throughput screening and discovery of new mutations; thus, it overcomes the obstacles of microarray-based genotyping technology. Accordingly, NGS allows comprehensive analysis of the frequency, evolution of alleles and allelic markers which enables better elucidation of complex blood group systems such as ABO.

5.2 Aim of the study

Following the feasibility of the NGS approach for genotyping of *FY* and *JK* blood group genes in previous chapters, the NGS approach together with long-range polymerase chain reaction (LR-PCR) will be used here for genotyping of the *ABO* blood group gene. The entire *ABO* gene will be sequenced here, including all 7 exons, introns and flanking regions. Previous genotyping studies only sequenced exons 6 and 7 (Lang et al., 2016) or left out the large intron 1 (Huh et al., 2011, Thuresson et al., 2012); in contrast, sequencing the entire gene here will enable a more comprehensive genotyping of *ABO*, which will allow identification of rare or novel polymorphisms (including intronic) with the resulting alleles and correlation of these polymorphisms with *ABO* alleles.

5.3 Results

5.3.1 LR-PCR of the *ABO* gene

A total of 47 samples were used for NGS genotyping of the entire *ABO* gene including flanking regions (Table 2.3). The ABO serological phenotype (A, B, O or AB) was provided for all samples and used as a basis for their selection (Table 5.1). Four primer pairs were designed using Primer3 and the NCBI database (NCBI, 2016c)(NCBI, 2016c) (see section 2.2.4.1, Table 2.3 for more details about the primer pairs), resulting in four overlapping amplicons covering the entire *ABO* gene and outer regions (Figure 5.1). For the amplification of the first *ABO* area (the upstream region, exon 1 and part of intron 1), two primer pairs (1A and 1B; section 2.2.4.1) were used. Initially, several samples were not successfully amplified by primer pair 1A, which were adapted from the study by Huh et al. (2011), likely due to polymorphisms around the primer binding site (section 5.3.4.2); therefore, primer pair 1B was designed to amplify the rest of samples. For LR-PCR amplification using these primer pairs (1A or 1B), the Phusion Flash High-Fidelity PCR Master Mix (Thermo Scientific, Leicestershire, UK) was used under thermocycling conditions adapted from Huh et al. (2011), as shown in Table 2.7 in Chapter 2. Amplicons 2, 3 and 4 were amplified using the LongAmp® Hot Start *Taq* Master Mix (New England BioLabs Inc, Herts, UK) under thermocycling conditions shown in Table 2.6 in Chapter 2. Figure 5.2 shows examples of PCR amplicons of *ABO*, which were loaded onto 1% agarose gels and separated by electrophoresis (see section 2.2.4.3). Amplicons were then purified using the magnetic bead technique (section 2.2.4.4) and subsequently quantified using the Qubit® 2.0 Fluorometer with Broad range (BR) assay Kit (InvitrogenTM, Paisley,UK) to ensure the 100 ng concentration required for the fragmentation process (section 2.2.4.5).

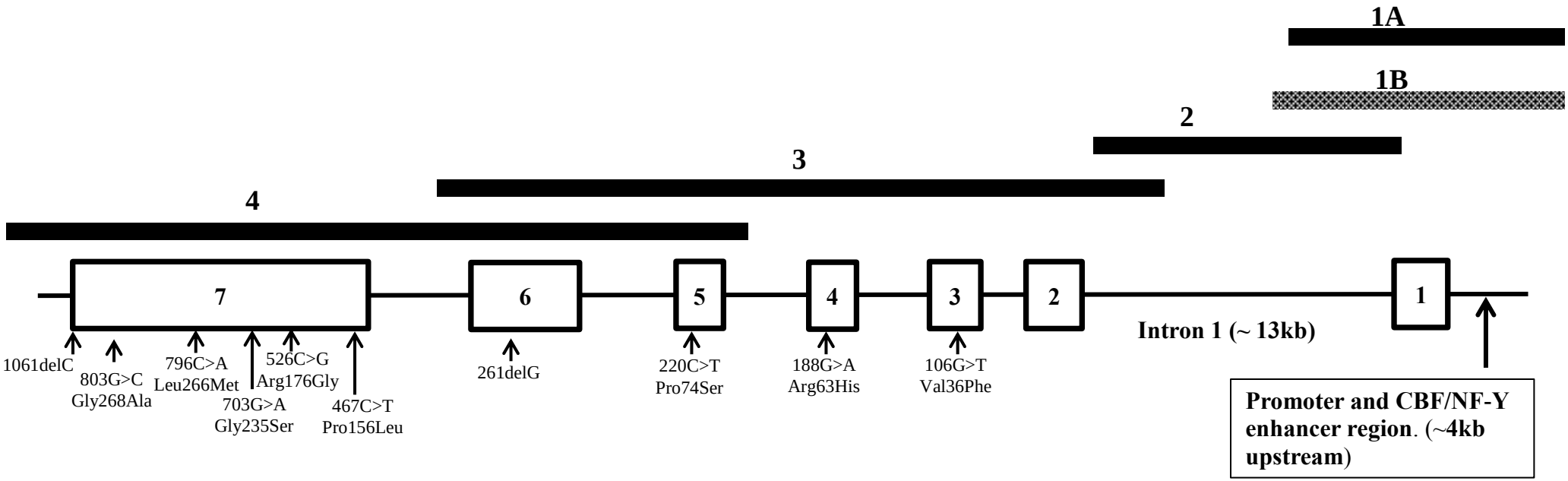


Figure 5.1 Four overlapping LR-PCR amplicons covering the entire *ABO* gene.

The *ABO* gene is approximately 20039 bp (according to NCBI GenBank) in size, comprising 7 exons (shown here as boxes) and 6 introns (shown here as a fine line). The overlapping amplicons 1A (6260bp) or 1B (7017bp), 2 (9188bp), 3 (10354bp) and 4 (9628bp) cover the entire gene plus the up-/down stream flanking regions. The key polymorphisms, most of which encode for a single amino acid change A^1 (*A101*) from *B* (*B101*) allele (Gly268Ala, Leu266Met, Gly235Ser, and Arg176Gly), (1061delC and Pro156Leu) are associated with A^2 (*A102*) allele, the 261delG are with O^1 (*O01*), which is also associated with (*O02*) allele along with Pro74Ser, Arg63His and Val36Phe. The exons are shown in reverse order here due to the anti-sense direction of the gene on chromosome 9. Refer to Table 2.3 for more detailed information on amplicon locations.

Table 5.1 The serology information of all 47 blood samples provided by the National Health Service Blood and Transplant (NHSBT; Filton, Bristol UK). * denotes the working sample ID. The **ABO001** number, which is the ISBT number assigned for the ABO blood group system (Daniels, 2013, Reid et al., 2012), is used here to name the samples accordingly for *ABO* NGS genotyping.

Sample number	Sample ID*	Phenotype	Sample number	Sample ID*	Phenotype	Sample number	Sample ID*	Phenotype
ABO001.01	3	A	**ABO001.21**	RN	O	**ABO001.41**	109	AB
ABO001.02	4	A	**ABO001.22**	R9	O	**ABO001.42**	111	AB
ABO001.03	7	A	**ABO001.23**	130	O	**ABO001.43**	112	AB
ABO001.04	25	A	**ABO001.24**	19	O	**ABO001.44**	114	AB
ABO001.05	40	A	**ABO001.25**	26	O	**ABO001.45**	115	AB
ABO001.06	73	A	**ABO001.26**	28	O	**ABO001.46**	118	AB
ABO001.07	44	A	**ABO001.27**	57	B	**ABO001.47**	119	AB
ABO001.08	64	A	**ABO001.28**	8	B			
ABO001.09	72	A	**ABO001.29**	60	B			
ABO001.10	RL	A	**ABO001.30**	32	B			
ABO001.11	RH	A	**ABO001.31**	56	B			
ABO001.12	11	O	**ABO001.32**	62	B			
ABO001.13	17	O	**ABO001.33**	67	B			
ABO001.14	50	O	**ABO001.34**	RP	B			
ABO001.15	66	O	**ABO001.35**	R7	B			
ABO001.16	10	O	**ABO001.36**	101	B			
ABO001.17	12	O	**ABO001.37**	103	B			
ABO001.18	29	O	**ABO001.38**	125	B			
ABO001.19	38	O	**ABO001.39**	131	B			
ABO001.20	RG	O	**ABO001.40**	105	AB			

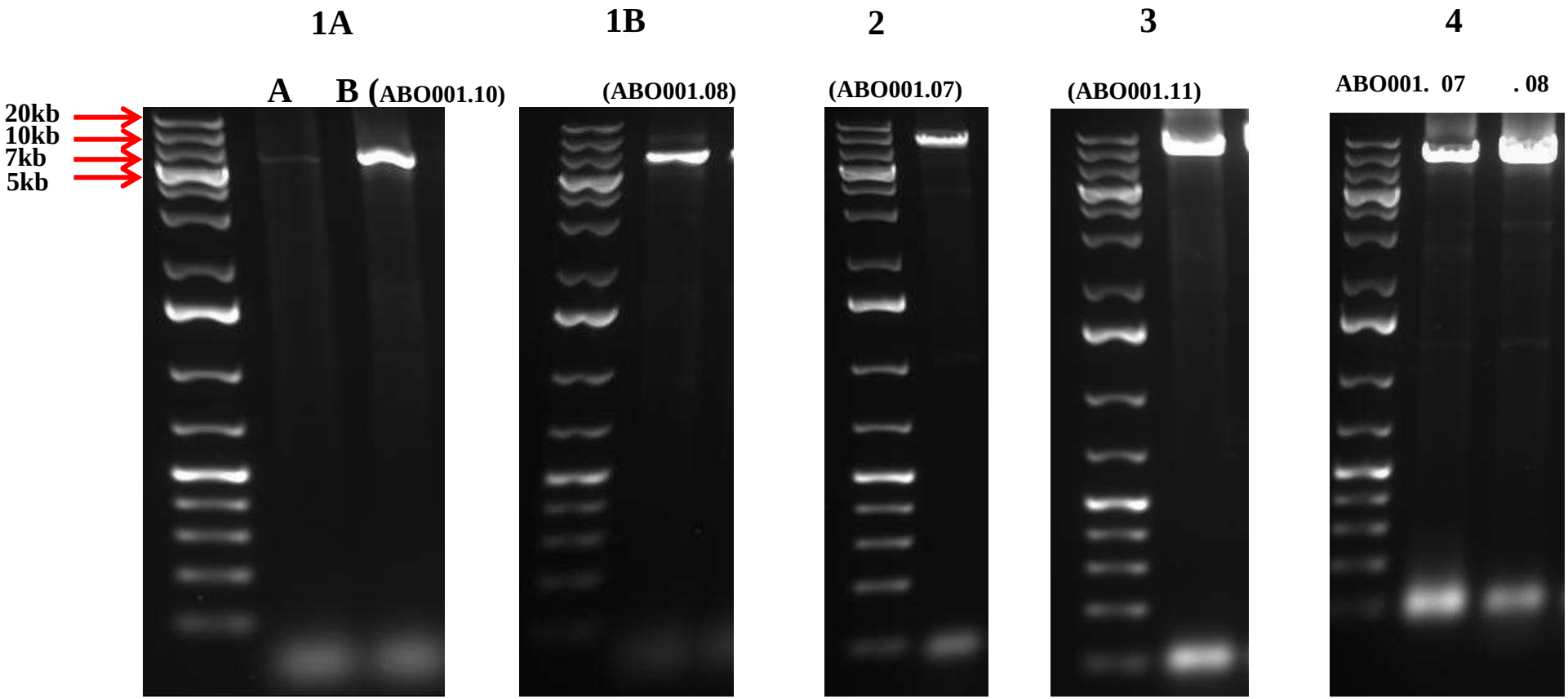


Figure 5.2 Four LR-PCR amplicons covering the whole *ABO* gene.

Four LR-PCR amplicons were used to amplify the entire *ABO* gene for each sample. The amplicons are the product of primer pairs 1A (6260bp) or 1B (7017bp), 2 (9188bp), 3 (10354bp) and 4 (9628bp). Here, examples of amplicons are shown aligned to the DNA marker. Amplicon 1A from sample A (the ID of the sample was R8, phenotype O) was unsuccessfully amplified, was not used in genotyping, but was successfully amplified from sample B(ABO001.10). Example of successful amplification of Amplicons 1B, 2 and 3 are shown. Successful amplification of 2 samples using primer pair 4 is shown. All amplicons were electrophoretically separated on 1% agarose gel at 80 V for 1 hour and 40 minutes.

5.3.2 NGS of the *ABO* gene

5.3.2.1 *ABO* purified fragmented library

Following amplicon purification and quantitation, the NGS sequencing libraries were fragmented by enzymatic shearing using the Ion Xpress™ Plus Fragment Library Kit (section 2.2.4.6). The 4 amplicon-pool (100ng) from each sample were enzymatically fragmented for 8 minutes, which enabled a wide fragment size distribution and sufficient yield (peak) around 200-300bp. Subsequently, samples were purified using SPRIselect® reagent kit (Beckman Coulter, UK) (section 2.2.4.6).

The outcome of the fragmentation (fragment size distribution) was assessed by the Agilent® 2100 Bioanalyzer and Agilent High Sensitivity DNA Kit (Agilent Technologies UK Limited). Figure 5.3 illustrates the size distribution outcome of the purified fragmented ABO sequencing library.

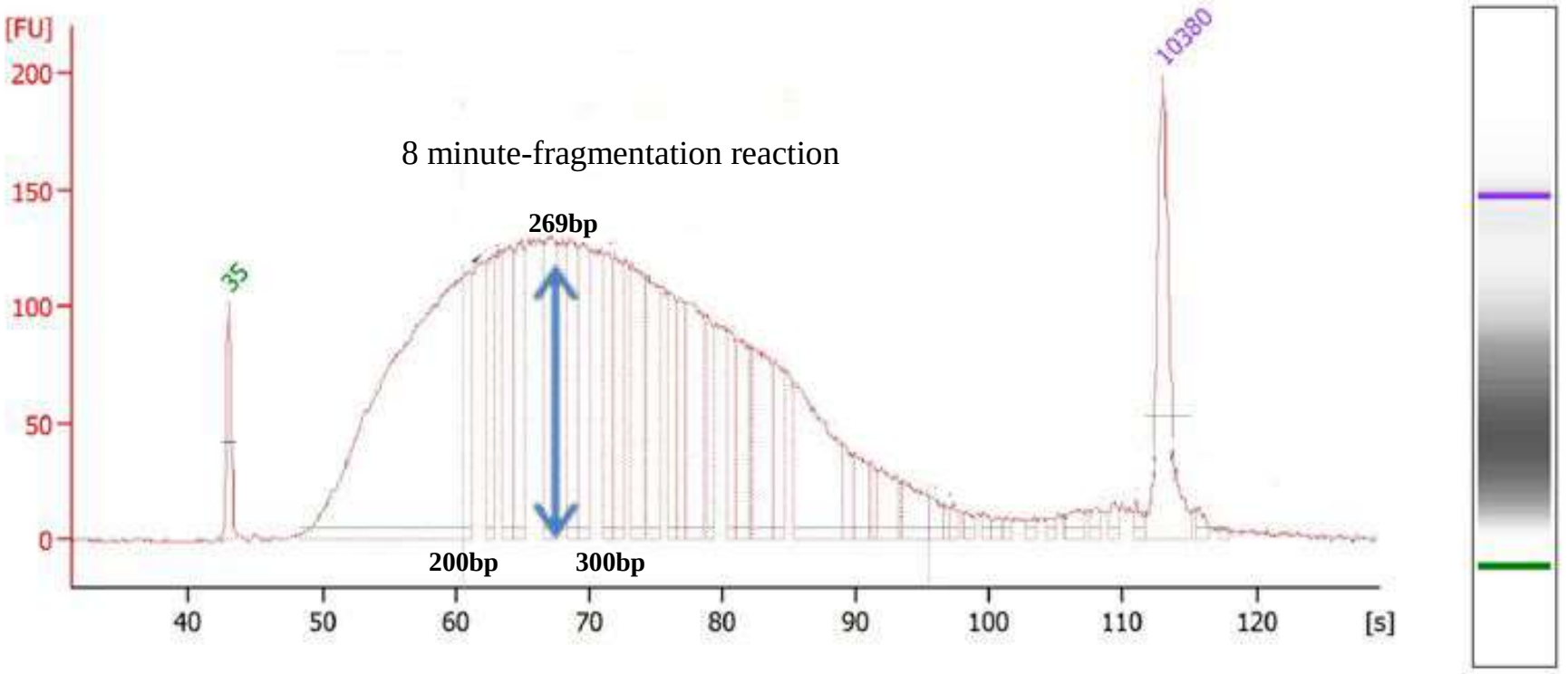


Figure 5.3 An electropherogram of a purified fragmented *ABO* DNA library (consists of a pool of four *ABO* amplicons).

An *ABO* sample was fragmented using the Ion Xpress™ Plus Fragment Library Kit for 8 minutes. A wide fragment size distribution can be seen with a peak around 200-300bp (marked by a blue arrow), which is recommended for the 200bp read length on the Ion PGM™. The green number (35bp) is the lower marker and the purple number (10380bp) is the upper marker. Results shown were obtained using the 2100 Bioanalyzer® instrument. The raw trace from the 2100 Bioanalyzer® instrument is shown in the appendix B.

5.3.2.2 Ligation of barcoded adapters (*ABO* purified-ligated library)

Following fragmentation and purification, samples were ligated with barcoded adapters (P1 and Ion Xpress™ Barcode X adapter, provided in the Ion Xpress™ Barcode adapters Kit), catalysed by DNA ligase. The ligated libraries were then purified using SPRIselect® reagent kit before assessing the ligated-purified libraries by Agilent® 2100 Bioanalyzer and the Agilent High Sensitivity DNA Kit (Agilent Technologies UK Limited) (section 2.2.4.9). Figure 5.4 illustrates the size distribution of an *ABO* sample sequencing library after ligation.

5.3.2.3 Size selection

The ligated libraries were then size-selected to provide a suitable size range (around 200bp) for the Ion PGM™ Template OT2 200 kit. In the process of sequencing 47 samples, 3 separate sequencing experiments were conducted, in all of which the SPRIselect® was used for size selection (section 2.2.4.10). Figure 5.5 shows an *ABO* sample that was size-selected by SPRIselect®. Next, the concentrations of the size-selected libraries were obtained and calculated using the 2100 Bioanalyzer® instrument (section 2.2.4.11) for the purpose of template preparation (section 2.2.5), prior to sequencing using the Ion Torrent PGM™ (Section 2.2.6).

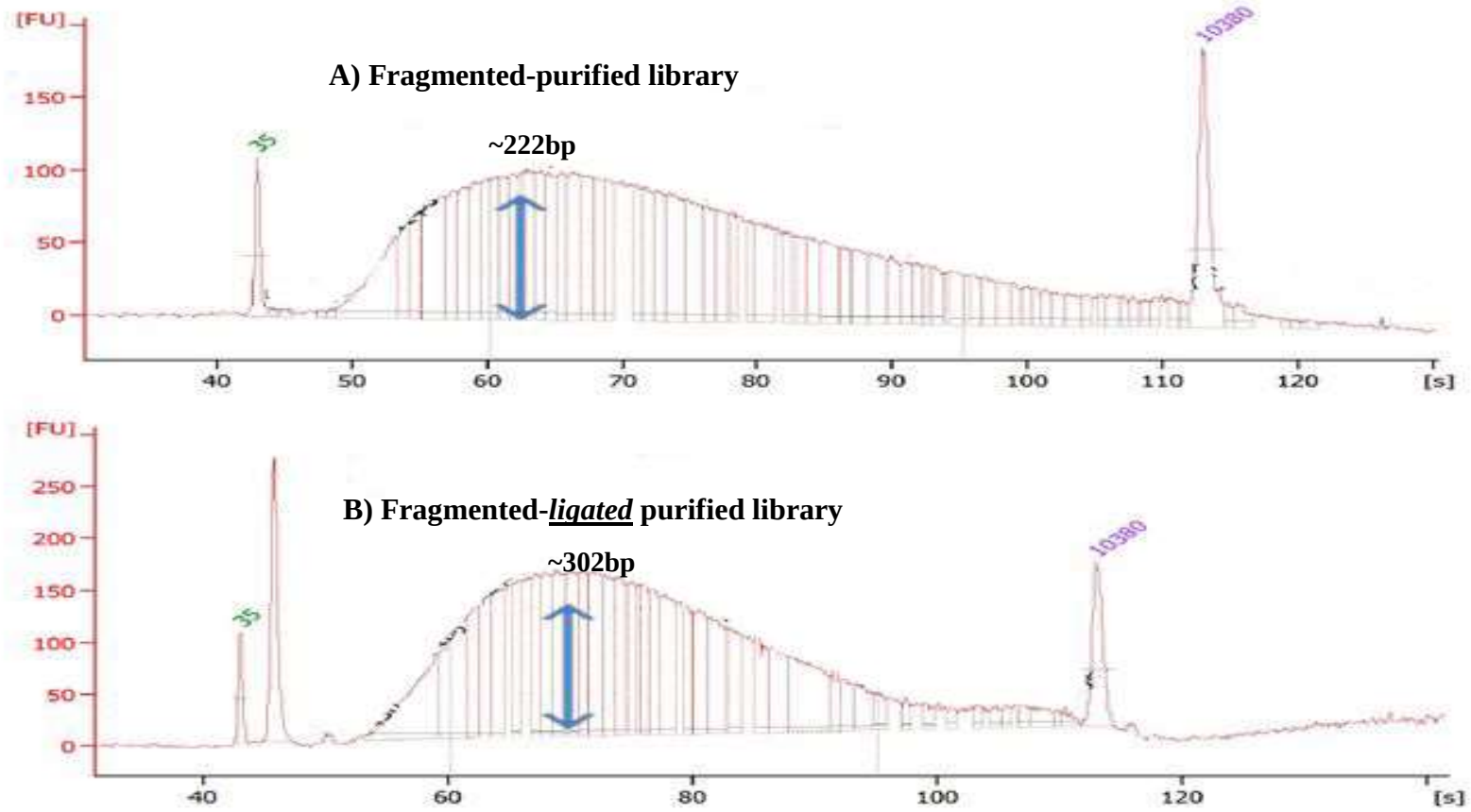


Figure 5.4 Two electropherograms of an *ABO* sample library.

A) Fragmented-purified library with a peak around 200-300bp.
B) Fragmented-ligated purified library with the peak shifted to the right (due to adapter ligation, represented by an increase in size).

The green number (35bp) is the lower marker and the purple number (10380bp) is the upper marker. Results shown were obtained using the 2100 Bioanalyzer® instrument. The raw trace from the 2100 Bioanalyzer® instrument is shown in the appendix B.

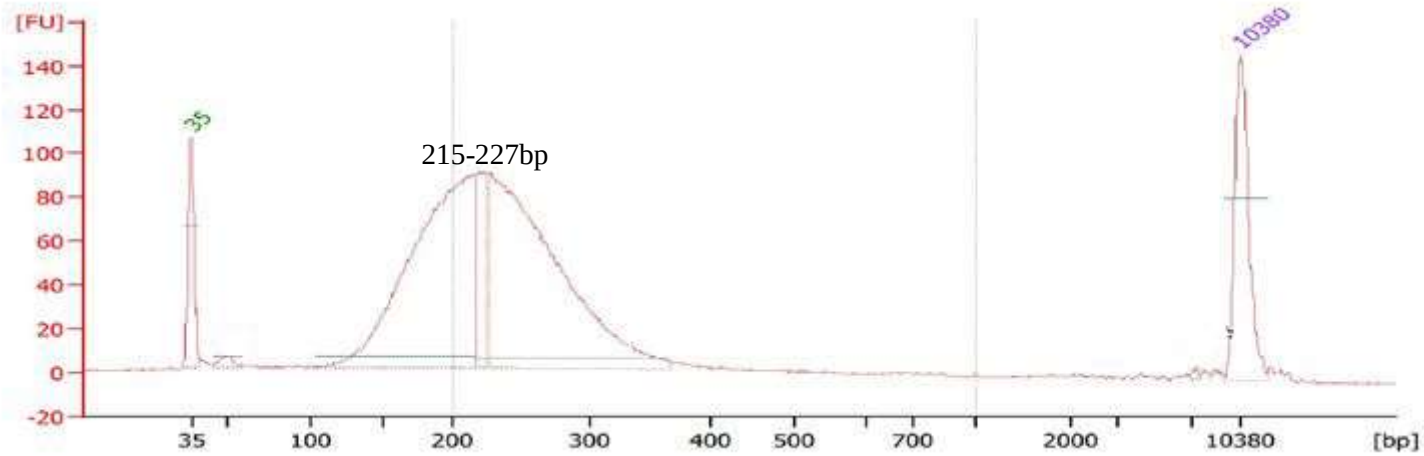


Figure 5.5 An electropherogram of a size-selected *ABO* sequencing library by SPRIselect®.

An *ABO* sequencing library was size-selected using SPRIselect® magnetic beads. A peak around 200 bp (215-227bp) was achieved. The green number (35bp) is the lower marker and the purple number (10380bp) is the upper marker. Results shown were obtained using the 2100 Bioanalyzer® instrument.

5.3.3 NGS data quality control

5.3.3.1 Sequencing data summary report

Sequencing of the 47 *ABO* samples was conducted in 3 separate experiments using Ion PGM™, the raw data of which was then processed by the Torrent suite server, to produce a summary run by Torrent Suite™ Software Version 4.4. The average total usable reads generated from the 47 *ABO* samples was 3,399,21, with a mean coverage depth of 650X. A summary of the sequencing report is shown in Table 5.2 and a representative *ABO* sequencing experiment brief report is depicted in Figure 5.6. In this sequencing report, the ISP loading density for the 316™ chip was 74%, with 26% of the wells remaining empty. The total number of reads usable for downstream analysis was ~3.7 million. These reads were obtained following well classification and made up 80% of the total reads containing library ISPs. Also, these ISPs underwent filtration to remove polyclonal ISPs (13%), test fragments (1%), adapter dimer (3%) and low quality ISPs (5%). The mean read length was 145bp (Figure 5.6).

Table 5.2 A summary of the Ion PGM™ sequence output for the 47 *ABO* samples, processed in 3 separate runs. *Samples were not genotyped due to issues, such as uncomplete sequencing coverage in areas of the gene.

	No. of samples loaded into the chip	No. of samples genotyped	No. of Samples not genotyped*	ISP Loading %	Total usable reads	Usable reads%	Mean read length
1st run report	16	11	5	79	3,504,607	70	134 bp
2nd run report	20	20	-	74	3,723,345	80	145 bp
3rd run report	21	16	5	85	2,970,010	55	141bp

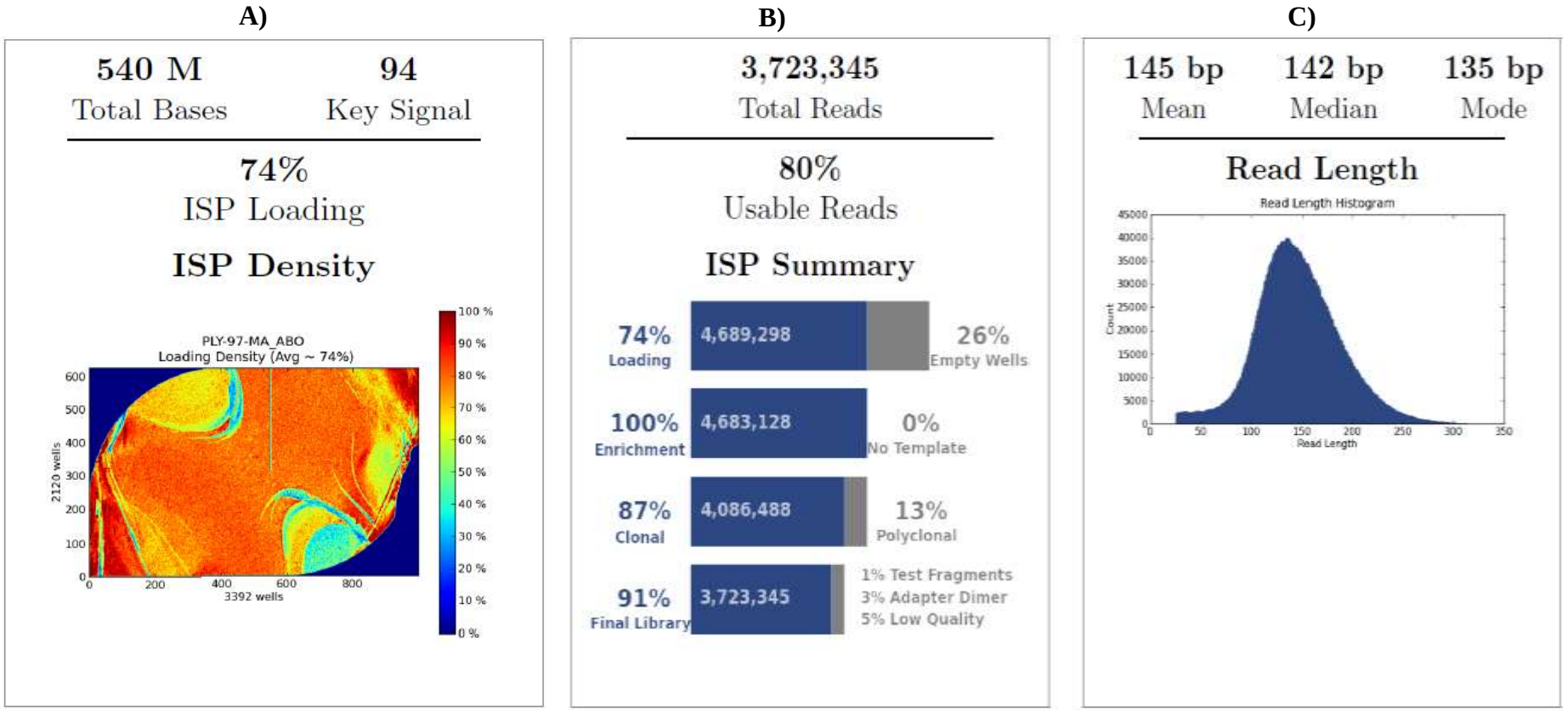


Figure 5.6 A summary report for a single *ABO* library sequencing run.

A) 74% of chip wells contained the ISPs. The colour represents the loading percentage of ISP across the 316™ chip plate surface.

B) The total number of usable reads is 3,723,345, which are provided after trimming and filtration from empty wells, non-templated, polyclonal reads and low quality ISPs. 80% of ISPs was obtained by dividing these reads by the total number reads containing the library ISPs (4,648,719). The live/enrichment percentage is 100%, which indicates that ISPs contain a strong sequence signal from the test fragment and template library.

C) A histogram shows a mean reading length of 145bp. The read count is displayed in the y-axis, while the read length, in bp, is shown on the x-axis.

5.3.3.2 NGS data quality control

The quality of NGS data is important for further analysis, which includes genotyping. The quality and the mean coverage depth (~650X) of libraries were analysed using the Ion Torrent Suite™ plugins: coverage analysis and FastQC. The mean coverage depth was high (650X), which increased the confidence for variant calling. The per base and per sequence quality were in turn assessed by FastQC according to the Phred score, which is a measure logarithmically related with the base calling error probability and is used to assess the quality of bases generated from automated sequencing (Ewing et al., 1998).

5.3.3.2.1 Per base sequence quality

Figure 5.7 demonstrates the *ABO* sequence quality (the base pairs sequenced) according to the Phred score. The mean quality score here was ~29-30, which then gradually decreased towards the end of sequencing; this phenomenon has been described before as a normal observation in high-throughput platforms (Andrews, 2016). This mean quality score range indicated a high base call accuracy (> 99% to 99.9%) with a probability of 1 in 1000 of incorrect base call. The quality of the samples in all three runs was comparable (above 99% base call accuracy, according the Phred score; Table 5.3).

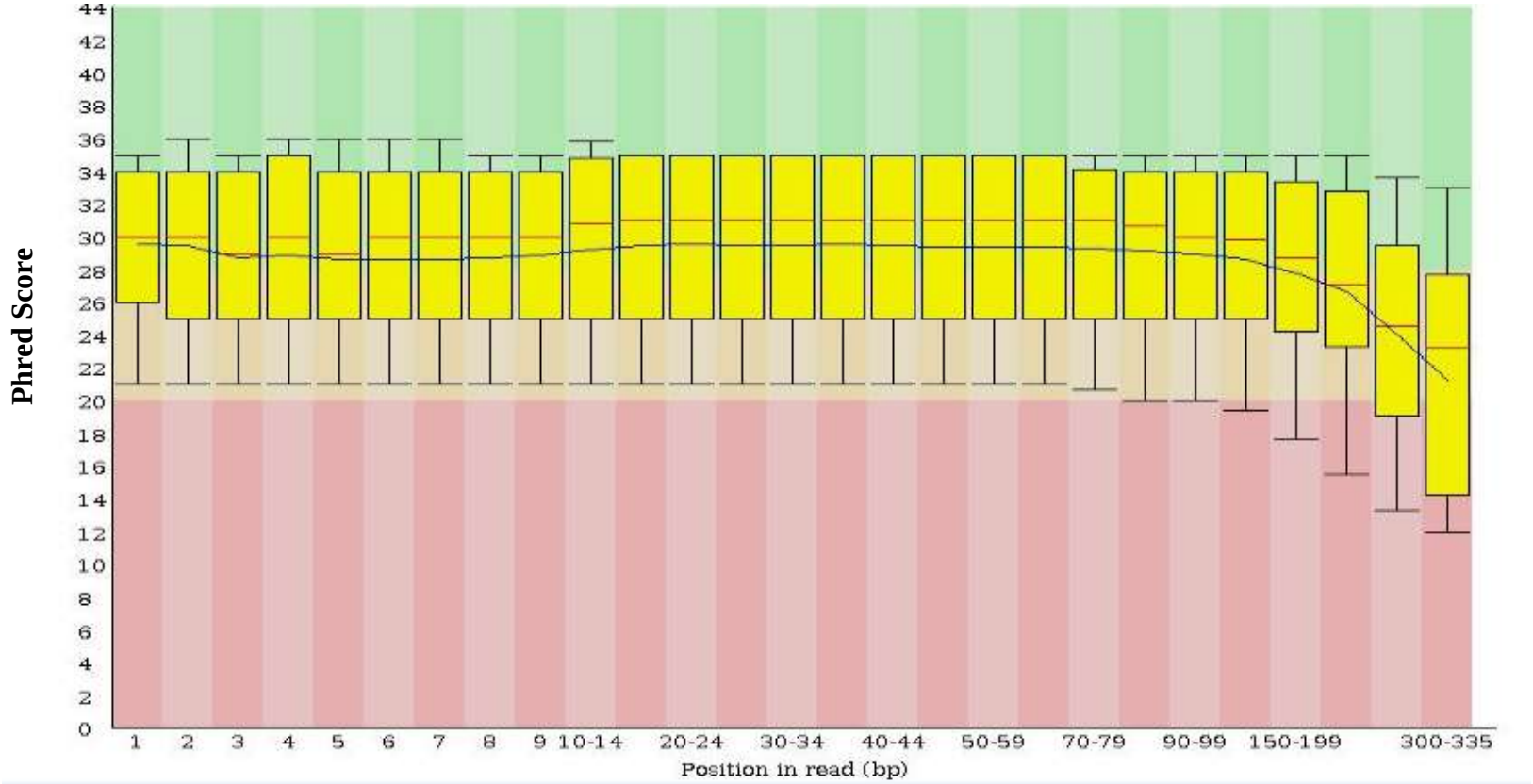


Figure 5.7 The Phred quality score across all bases of a single representative *ABO* sample.

The x-axis represents the position of bases in the read, while the Phred score is displayed on the y-axis. The three-colour background divides the y-axis into three regions of different quality levels, according the Phred score: very good (green), reasonable (orange) and poor quality base calling (red). The yellow boxes represent a BoxWhisker-type plot for each position with a 25-75% interquartile range. Upper and lower whiskers represent the 10% and 90% points. The blue line represents the mean value of the base call quality (29-30), which indicates ~ 99.9% base call accuracy. The median value of the quality is denoted by a red line. All *ABO* samples showed comparable output.

5.3.3.2.2 Per sequence quality score

The quality of the sequences was also assessed using FastQC (see Figure 5.8 for an example). Here, the mean sequence quality of *ABO* was high, with most sequence reads achieving a Phred quality score of 30, in accordance with the previous section (5.3.3.2.1). The quality of all three *ABO* runs was comparable (above 99% base call accuracy, according to the Phred score; Table 5.3). As a result, these high quality parameters, along with the coverage depth, provide confidence in the *ABO* NGS data for further analysis.

Table 5.3 Summary of the sequencing quality of the three *ABO* NGS experiments. All runs achieved a base call accuracy of ~ 99.9%.

Experiment	Per base sequence quality	Per sequencing quality score
1st	29-30	29-30
2nd	29-30	30
3rd	29-30	30
Base call accuracy	~99.9%	~ 99.9%

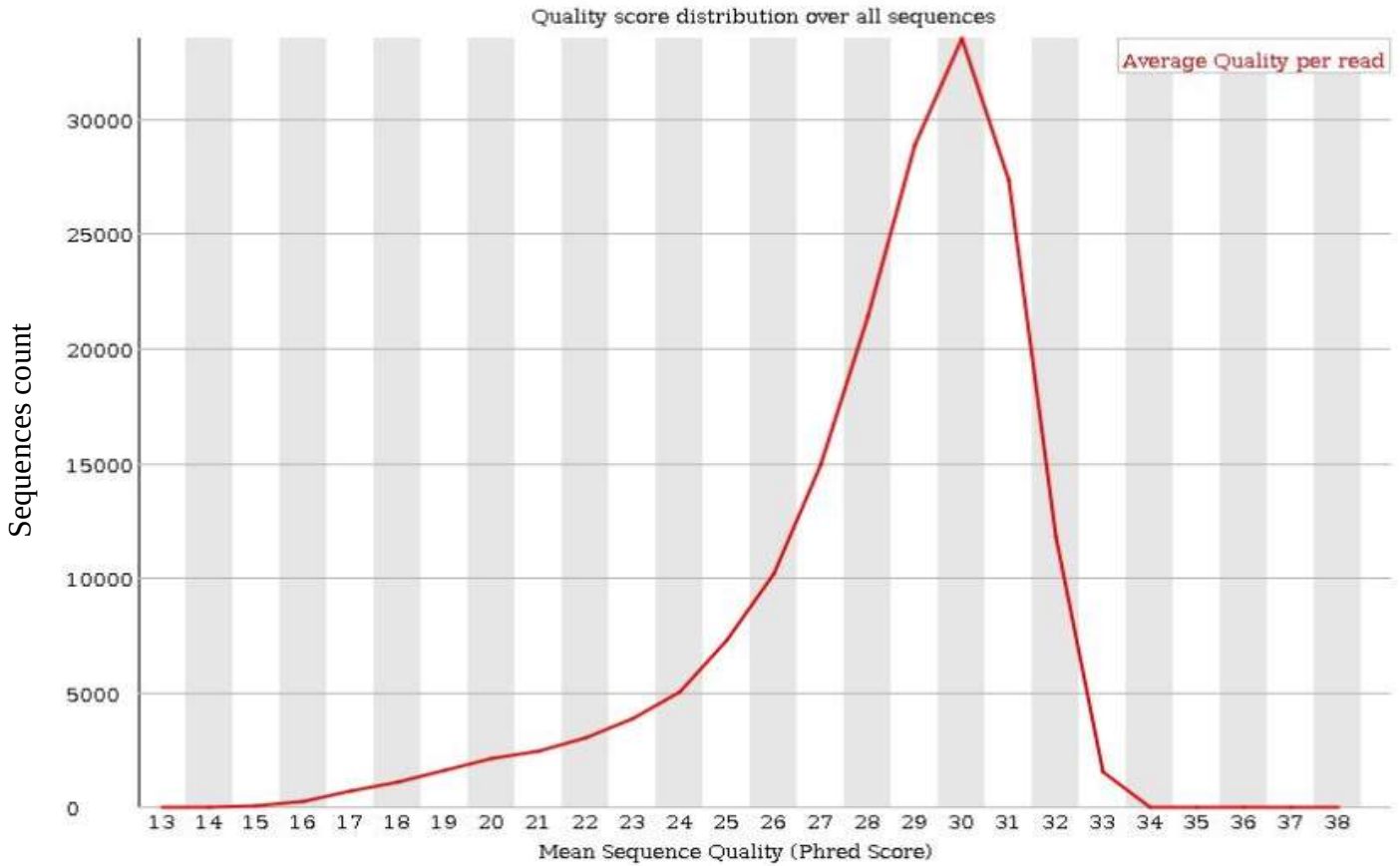


Figure 5.8 The mean quality score of *ABO* sequences generated from a single sample.

The mean quality score (Phred score, x-axis) of the sequences across the number of reads (y-axis) is shown. The mean quality score for most of the sequences was 30 **(**Phred score**)**, which indicates high quality of generated reads with an accuracy of 99.9% and a probability of 1 in 1000 that the base was incorrectly called. All *ABO* samples showed comparable output.

5.3.4 NGS sequence visualisation

The IGV software (version 2.3) was used to visualise the *ABO* sequencing data, which was aligned to the *ABO* reference sequence in the human genome (hg19) and associated with the mRNA accession number (NM_020469.2). Different aspects were assessed using this tool: the integrity of the sequence (full coverage across the *ABO* gene) and the coverage depth. Moreover, other obtained genomic information included mutations, such as SNPs, deletions, insertions, as well as the zygosity and chromosomal location of mutations, in addition to the correlation with amino acid locations. This information, together with genotyping software, such as Ion Reporter™, enables a comprehensive and efficient analysis of *ABO*. Figure 5.9 illustrates an *ABO* sample (phenotyped as A, genotyped as *A201(ABO*A2.01)/O75*) sequence of the entire gene and flanking regions, which reflects the integrity of the library preparation and the primer specificity.

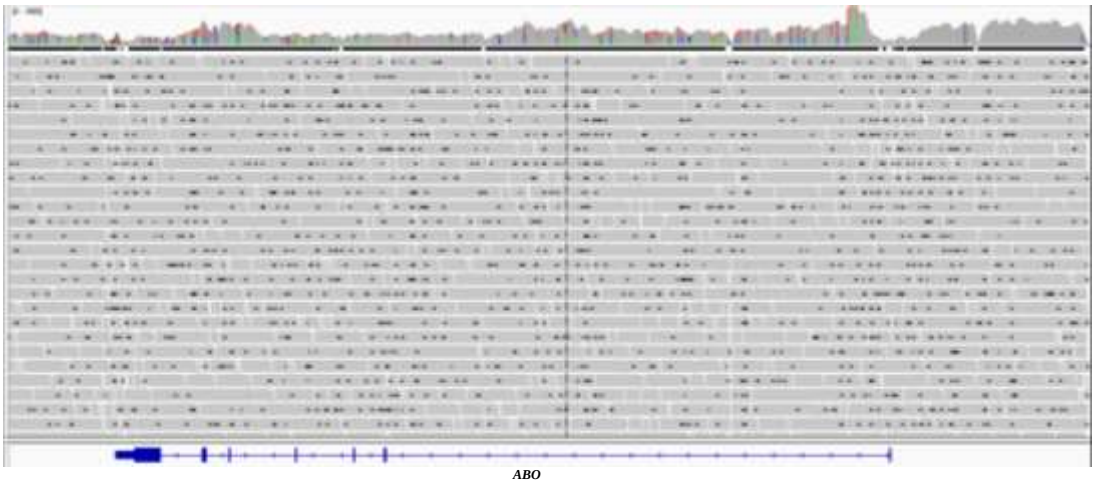


Figure 5.9 An IGV image illustrating the full coverage of the sequencing data from one *ABO* sample.

The reference *ABO* gene is fully covered and aligned by the sequencing reads. The coverage depth is shown on top (which appears in the software when the arraow is pointed at it). Coloured bars represent SNPs across the gene. The complete coverage enables the analysis of all polymorphisms across the gene, including those in critical areas such as splice sites and regulatory regions.

5.3.4.1 Observations on the *ABO* reference sequence

The analysis of the A*BO* sequence data, including the visualisation, revealed a number of observations with regards to the *ABO* reference sequence. The *ABO* reference gene sequence is displayed from right to left (an anti-sense direction), which should be taken into account when analysing SNPs. Figure 5.10 is an example issue the anti-strand *ABO* reference gene sequence poses when analysing the consensus 467C>T missense mutation, which encodes the amino acid change Pro156Leu. In the IGV, the SNP is observed as G>A instead of C>T, as a result of the anti-sense direction of the *ABO* sequence, which had to be flipped to reveal the reference nucleotide (C). Consequently, the mutation codon needs to read in the opposite direction (right to left) in the reverse-complement sequence. This issue was overcome manually (Figure 5.10) and the genotyping software Ion Reporter^{TM} provided a neutralised annotation (section 5.3.5). It was also noticed that the *ABO* reference sequence derived from the human genome (hg19) was assembled from regions of two *O* allele sequences (contigs). Exons 1 to 5 seemed to correspond to the *ABO*O.01.02* (*O02/O^{1v}*) allele as the amino acids Phe36, His63 and Ser74 were encoded, in contrast to the suggested consensus allele *ABO*A1.01* (*A101*) encoding for Val36, Arg63 and Pro74. On the other hand, exons 6 and 7 represent the sequence of allele *ABO*O.01.01* (*O01/O^{1}*), as there was a (261G) deletion in exon 6 of the reference sequence from hg19, while exon 7 was identical to that of the *ABO*A1.01* allele. Consequently, an insertion of G at this location (chr9:136132908-136132909) was noticed in samples carrying the *A* and *B* alleles, while samples with the *O01* allele showed no change from the reference sequence. The same was seen with the *ABO* reference sequence from the human genome reference sequence (hg38), where there was a 261G deletion in exon 6. Table 5.4 lists the differences between the *ABO* reference sequence and sequences of hg19 and the consensus *ABO*A1.01* (*A101*) RefSeqGene, NG_006669.1. Figure 5.11 shows an

example of an ABO sample with 261G insertion.

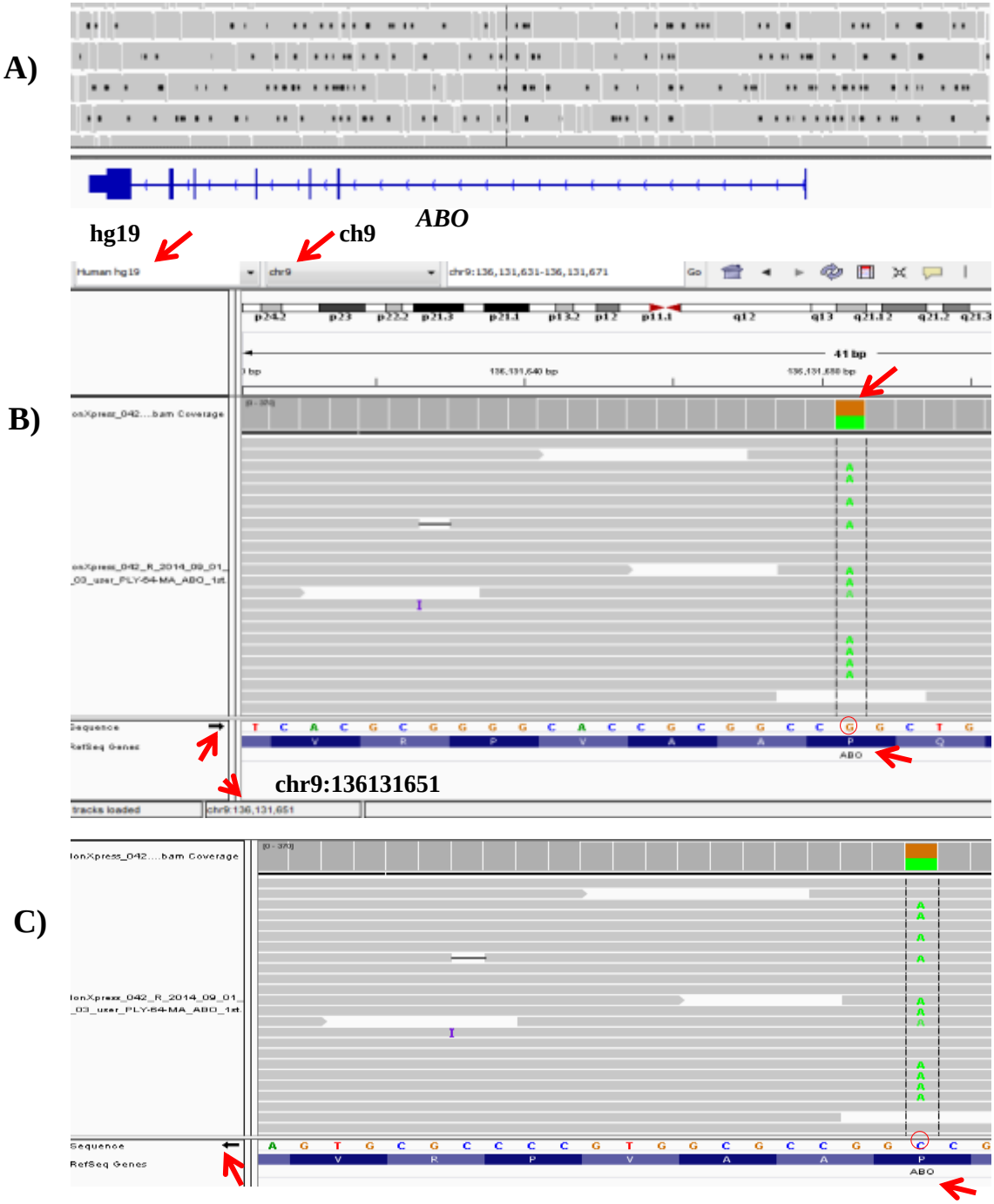


Figure 5.10 IGV visualisation analysis of *ABO* sequencing data of a single sample.

A) The orientation of the *ABO* reference gene is right to left (against the direction of the chromosome). Exon 7 is shown on the left through to exon 1 on the right.

B) A heterozygous SNP at the chromosomal location chr9:136131651 is shown as 467G>A, whereas the consensus is 467C>T. This is because the *ABO* reference sequence is illustrated as anti-sense strand.

C) The reference sequence was intentionally complemented to reveal the nucleotide in the sense strand, the consensus i.e. nucleotide C. The mutation is C>T but the sequencing data cannot be visually altered so the T cannot be shown, the mutation still looks like A. Note: the codon sequence should also be read from right to left (CCG Pro) instead of (GCC Ala). The SNP is heterozygous (split bar 50% C and T) 467C>T, encoding for Pro165Leu.

A)

```
Query  136132903  AGGGGT-ACCACGAGGACATCCTTCCTACTGCACATGGAGAGAGGCGTGCGGTCACATGG  136132961
Sbjct  17733      AGGGGTCACCACGAGGACATCCTTCCTACTGCACATGGAGAGAGGCGTGCGGTCACATGG  17674
```

B)

```
Query  133257516  AGGGGT-ACCACGAGGACATCCTTCCTACTGCACATGGAGAGAGGCGTGCGGTCACATGG  133257574
Sbjct  17733      AGGGGTCACCACGAGGACATCCTTCCTACTGCACATGGAGAGAGGCGTGCGGTCACATGG  17674
```

C)

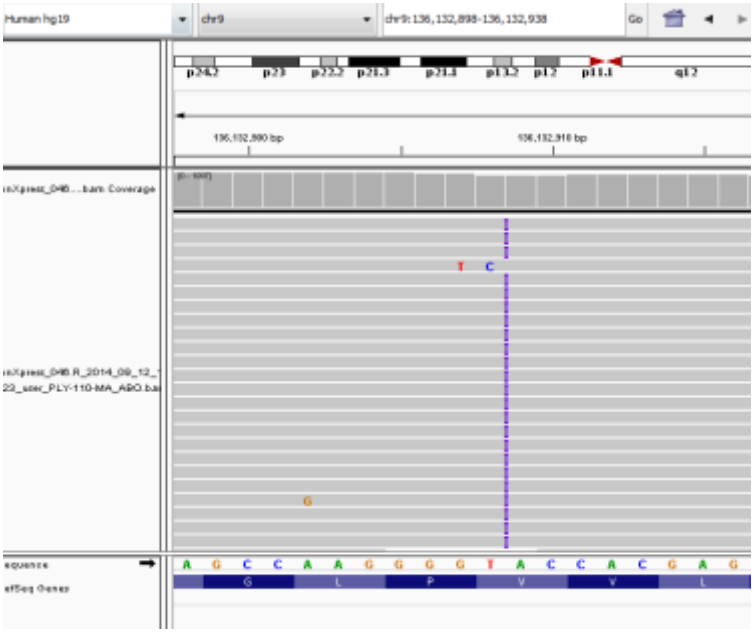


Figure 5.11 illustrations regarding the *ABO* reference sequence.

A) an alignment of the *ABO* reference sequence from hg19 (Query, NC_000009.11: 136130563-136150630) and the (sbjct, NG_006669.1) consensus NCBI own reference sequence (*ABO*A1.01*) revealed a G deletion (shown as a C here, due to anti-sense direction) in the query, represented by the red arrow.

B) A G deletion in the *ABO* reference sequence from the hg38 (query, NC_000009.12: 133255176-133275214) was observed when aligned with the NCBI reference sequence (RefSeqGene, NG_006669.1).

The alignment was conducted using the NCBI Blast tool.

C) An IGV visualisation analysis of an *ABO* sample (with serological phenotype AB) shows homozygous insertion of G (denoted by purple bars) at the chromosomal location chr9:136132908-136132909. The insertion is noticed as the reference sequence lacks the G at this location, which represents the sequence of *O01* allele. Conventionally, the 261Gdeletion *(O01)* leads to a frame shift that alters the amino acid sequence from 88 onwards and results in a premature stop codon after amino acid 117 (118 stop codon).

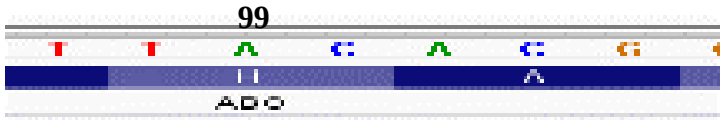
Table 5.4 The differences in polymorphisms between the reference sequence in hg19 and the agreed consensus sequence (*ABO*A1.01*) from the NCBI.

A) In exon 3, the amino acid Phe36 is seen in the hg19 ref sequence, His63 in exon 4 and Ser74 in exon 5. The opposite was found in the consensus sequence from NCBI (RefSeqGene, NG_006669.1). *Nucleotide 261G was found in exon 6 of the consensus sequence, but was absent in the reference sequence of both hg 19 and 38 (thus, an insertion is seen when aligning samples carrying *A* and *B* alleles. B) The reference sequences from hg 19 and 38 show His instead of Thr at location 99 due to single nucleotide shift to the right as a result of the G deletion in the hg19 reference sequence (ACA/Thr to TAC/His). *** The location of amino acids were slightly shifted in exon 7 but were neutralised using the Seattle online tool. In addition, the nucleotide numbers were changed by one nucleotide in exon 7. As a result, the reference sequence in hg 19 was suggested to be mixed between *O02(ABO*O.01.02)* (exon 1-5) and *O01(ABO*O.01.01)* (exon 6-7), while that in hg 38 was similar to that of *O01(ABO*O.01.01)*. The bold highlight polymorphisms that differe.

A)

	Exon 2	Exon 3	Exon 4	Exon 5	Exon 6	Exon 7
A. **Consensus ref Sequence**	Arg18Leu 53G>T	**Val 36Phe** 106G>T, Gly35Arg 103G>A	**Arg63His 188G>A, 189C>T**	**Pro74Ser 220C>T**	**Thr99 297A>G, 88fs 118 stop 261delG**	Pro354fs/1061delC, Leu310 930G>A, Tyr309Cys 926A>G, Val277Met/**829G>A**, Gly268Ala **803G>C**, Gly268Arg 802G>A, Leu266Met 796C>A, Pro257 771C>T, Ile256 768C>A, Gly235Ser **703G>A**, Pro227 681G>A, His219 657C>T, Phe216Ile 646T>A, Arg199Cys 595C>T, Trp181stop 542G>A, Arg176Gly 526C>G, Pro156Leu 467C>T, Arg161His 482G>A
B. Hg19 reference	Arg18Leu 53G>T	**Phe36Val** 106T>G Gly35Arg 103G>A	**His63Arg 188A>G, 189T>C**	**Ser74Pro 220T>C**	**Pro88fs 118 stop 261insG* His99Arg 296A>G ****	• **The amino acid locations slightly shifted, although neutralised***** • **All nucleotide positions minus 1, e.g.** Val277Met/**828G>A**, Gly268Ala **802G>C**, Gly235Ser **702G>A**.

**B)



5.3.4.2 Observations on the *ABO* amplicons

Four primer pairs (4 amplicons) were used to amplify the *ABO* gene; however, amplicon 1A, which covered the region ~5kb upstream of and including exon 1, and part of intron 1, was not amplified in a number of samples of different phenotypes and thus alleles (Figure 5.12). As a result, another amplicon (primer pair/1B) was designed in order to overcome this issue, which lead to successful subsequent amplification of samples processed in the runs after the 1st. In the 1st run, all samples were sequenced by NGS, including those not amplified by primer pair 1A, to establish why there was failure to amplify in certain samples. The IGV visualisation investigation revealed SNPs around the 1A primer binding site, which may have accounted for the amplification failure. Table 5.5 lists SNPs from samples of different *ABO* genotype/phenotype, some of which were not amplified by 1A. *ABO* samples carrying the *A*, *B* alleles revealed the SNP A>G (ch9:136149500) at the binding site of the 13th nucleotide (T) of the reverse primer. In addition, one B sample, one AB sample and O phenotype samples also displayed a SNP, T>A located at ch9:136155720, at the first nucleotide adjacent to the 3' end of the forward primer (Table 5.5). With regard to amplicon 2, at least five samples of AB and 1 sample O phenotype were not successfully amplified, despite there being no visible factor such as an SNP at the primer binding region (data not shown).

Table 5.5 The SNPs found around the 1A primer pair binding sites in examples of *ABO* samples with different phenotype.

Examples of *ABO* samples that showed unsuccessful PCR amplification of the amplicon 1A and carry SNPs that may have hindered amplification. SNP 136149500 (A>G) was found in samples of phenotype/genotype A, B and AB, while SNP 136155720 (T>A), at the 1st nucleotide adjacent to the 3' end of the reverse primer was found in samples of B, AB and O phenotype. * 3/4 B and 2/2 A samples were NGS sequenced successfully with primer pair 1B. Het, heterozygous. Hom, homozygous. ** This may not be accurate as the coverage was low at this SNP.

	Phenotype/ genotype	No. of samples	**136149500 A>G**	**136155720 T>A**
No amplification	**B*/(*B101/O01*)***	4	4 Het	1 Het
	A*/(*A102/O02*) and (*A201/A201*)*	2	1 Het, 1 Hom	none
	AB	1	1 Het	1 Hom**
	O	2	none	2 Hom
Amplified	**B**	5	None	none
	A	3	None	none
	O	5	None	none

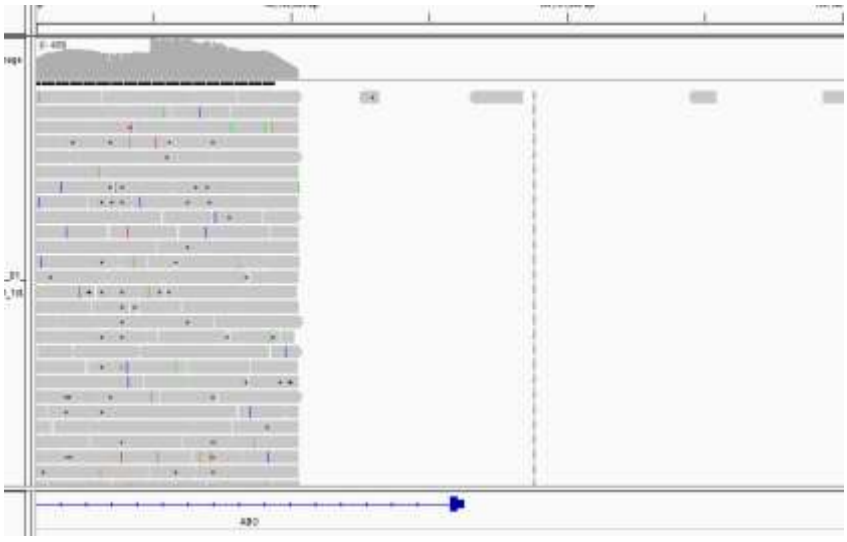


Figure 5.12 IGV image of an *ABO* sample of phenotype O where coverage of areas by amplicon 1A failed.

Part of intron 1- exon 1 and upstream area covered by the 1A primer pair showed no coverage of the sequence. This is due to failed PCR amplification, likely as a result of SNPs present at the primer binding sites.

5.3.5 Variant analysis and genotyping

The sequencing data generated from the Ion PGM™ was analysed using a combination of tools to enable accurate variant annotation. The Ion Suite™ plugin Variant Caller was used for mutation annotation, analysis of zygosity and production of VCF files, which were used in other variant annotation software for confirmation. The VCF files were uploaded to the online annotation online platform, SeattleSeq Annotation 137, which provided a wide range of information based on the NCBI database, including mutation chromosomal location, type of mutation, the number of nucleotides and the amino acid change and the rsID for a known mutation in the NCBI dbSNP database. Most of the variant analysis was conducted using the Ion Reporter™ 5.0 software (which only supports hg19 so far) to which the VCF files were also uploaded. This software provides extensive information, such as mutation type (single nucleotide variant/SNP or indel), mutation effect (missense or synonymous), mutation location (exon or intron), name of the gene, chromosomal location and rsID in dbSNP. The files can also be converted into different formats, such as PDF or Excel sheets. Moreover, this software overcame the aforementioned issues regarding anti-sense direction SNPs visualised using the IGV and provided the consensus SNP based on the NCBI database, which greatly simplified variant analysis. Regarding the issue of the nucleotide and amino acids shift arising from a shift in the *O* allele reference, the rsID and chromosomal location of the mutation was used to obtain the consensus order from the NCBI database, which was for genotyping of *ABO* samples in this study. Figure 5.13 shows part of the report from analysis of a sample with B phenotype and *B101(ABO*B.01)/O01(ABO*O.01.01*) genotype, generated by the the Ion Reporter™ 5.0 software.

A)

chr9:136131188	SNV	C	1	C/T
chr9:136131315	SNV	C	1	C/G
chr9:136131322	SNV	G	1	G/T
chr9:136131415	SNV	C	1	C/T
chr9:136131461	SNV	G	1	G/A
chr9:136131592	SNV	G	1	G/C
chr9:136131846	SNV	T	1	T/C
chr9:136131895	SNV	C	1	C/T
chr9:136131905	SNV	T	1	T/C
chr9:136132010	SNV	T	1	T/C
chr9:136132012	SNV	T	1	T/C
chr9:136132168	SNV	T	1	T/C
chr9:136132350	SNV	T	1	T/C
chr9:136132516	SNV	G	1	G/A
chr9:136132525	SNV	T	1	T/C
chr9:136132617	SNV	G	1	G/A
chr9:136132633	SNV	A	1	A/G
chr9:136132754	SNV	C	1	C/A
chr9:136132873	SNV	T	1	T/C
chr9:136132908	INDEL	T	1	T/TC

B)

ABO	NM_020469.2	exonic	synonymous	CTA	7	p.(=)	c.929G>A
ABO	NM_020469.2	exonic	missense	GCG	7	p.Gly267Ala	c.802G>C
ABO	NM_020469.2	exonic	missense	ATG	7	p.Leu265Met	c.795C>A
ABO	NM_020469.2	exonic	missense	AGC	7	p.Gly234Ser	c.702G>A
ABO	NM_020469.2	exonic	synonymous	CAT	7	p.(=)	c.656C>T
ABO	NM_020469.2	exonic	missense	GGC	7	p.Arg175Gly	c.525C>G
ABO	NM_020469.2	intronic					
ABO	NM_020469.2	intronic					
ABO	NM_020469.2	intronic					
ABO	NM_020469.2	intronic					
ABO	NM_020469.2	intronic					
ABO	NM_020469.2	intronic					
ABO	NM_020469.2	intronic					
ABO	NM_020469.2	intronic					
ABO	NM_020469.2	intronic					
ABO	NM_020469.2	intronic					
ABO	NM_020469.2	intronic					
ABO	NM_020469.2	intronic					
ABO	NM_020469.2	intronic					
ABO	NM_020469.2	exonic	missense	CGT	6	p.His99Arg	c.296A>G
ABO	NM_020469.2	exonic	frameshiftInsertion	ACC	6	p.Pro88fs	c.260_261insG

Figure 5.13 A section of the report generated by the Ion Reporter™ showing genotyping analysis of an *ABO* sample with B phenotype and *B101(ABO*B.01)/ O01(ABO*O.01.01)* genotype.

The highlighted polymorphism is used as an example

A) The chromosomal location of each variant, the type of mutation, reference and allelic nucleotide and zygosity are shown. The SNP or SNV at chr9:136131322 is shown as G>T heterozygous, which is different from the consensus SNP in the NCBI database.

B) Gene name, reference number, location of the mutation (exonic or intronic), the mutation effect, exon number and amino acid change are shown. Note that the SNP (marked) has been converted to the consensus SNP C>A.

5.3.6 *ABO* genotyping results from NGS

Extensive genotyping of the *ABO* gene using NGS allows identification of known common and rare mutations and discovery of novel mutations, which offers significant advantage over array based methods. The genotyping results and the interpretations of the SNPs and the amino acid changes were conducted according to the consensus reference sequence in the NCBI database (as if the NGS data were aligned against *A101 (ABO*A1.01)*) in order to be in line with conventional analysis. The results from 47 sequenced samples revealed various mutations, including SNPs and indels, both in exons (one novel) and introns of *ABO*. The genotypes of samples were then correlated with the phenotypes determined by serology (NHSBT). It is worth noting that the terminology used to describe *ABO* alleles here will be mainly the same as that used in dbRBC (Blood Group Antigen Gene Mutation Database), only for simplification reasons, for example *A101* for *ABO*A1.01* ISBT name, although in some cases the latter will be used for clarification.

5.3.6.1 *ABO* mutations in exons

NGS data revealed several mutations in exons of different *ABO* alleles, including rare and novel SNPs, in various ABO phenotypes (A, B, O and AB) (Table 5.6). The provided phenotypes of the samples sequenced were as follows: 11 A, 13 B, 15 O and 8 AB. The NGS genotyping data matched these phenotypes and, in addition, offered more detail on predicted A and A_2 phenotype.

Regarding samples of A phenotype, 9/11 samples were heterozygous with *O* alleles (Table 5.6 A). The common *A101* allele was heterozygous with *O01* (*A101*/*O01)* in one sample and with *O26* in two samples; the latter allele is similar to *O01* but features a silent mutation (768C>A) (Hosseini-Maaf et al., 2005, Roubinet et al., 2001). Two

samples carried the *A101* allele together with the common *O02* allele, which carries nine SNPs across exons 3, 4, 5, 6 and 7, in addition to the 261 G deletion (Reid et al., 2012) (Table 5.6). An allelic combination of *A101*/*O063* was found in one sample; the *O63* allele is rare, according to dbRBC (BGMUT), and contains a unique SNP 103G>A (the frequency of which is 0.02 % in Europe), according to the NCBI database and the 1000 genomes (NCBI, 2016a). This allele is listed in the Antigen Gene Mutation Database (BGMUT, dbRBC) by Stabentheiner et al. in an unpublished report as stated in BGMUT). The molecular basis of the *O63* allele is similar to that of the *O02* allele, except with the addition of a missense 103G>A SNP encoding amino acid change Gly35Arg. One sample was found to carry the *A102* allele with the 467C>T SNP (Pro156Leu), which is common among Asians (Reid et al., 2012). Four samples carried the *A201* allele (1 homozygous and 3 heterozygous), which differs from the *A101* allele by a 467C>T SNP (Pro156Leu) in exon 7 and a 1061C deletion codon directly adjacent to the translation stop codon (TGA). Consequently, the gene product of this mutation (GTA) has an extra 21 amino acids at the C-terminus, reflecting disruption of the stop codon (Yamamoto et al., 1992). Three samples were heterozygous for *A201*; one carried the *A101* allele and two carried *O75* and *O03*, respectively. The *O03* allele lacks the 216G deletion but contains 5 SNPs, 4 of which lead to amino acid changes in exon 7 (Gly268Arg and Arg176Gly) and exons 5 and 2 (Pro74Ser and Arg18Leu, respectively) (Reid et al., 2012). The *O75* allele is similar to the *O02* allele, with the addition of a rare SNP (542 G>A; 0.02% frequency in Europe, according to NCBI/1000 genome) which changes Trp181 into a stop codon. As a result, the suggested phenotypes of A samples according to the genotyping were: 8 A_1 and 3 A_2.

With regard to the samples of O phenotype, several *O* alleles were found. The *O01* allele was homozygous in 6/15 samples and heterozygous in 3 samples; the *O02* allele was homozygous in one sample and heterozygous in one sample; the *O03* allele was

heterozygous in two samples; and the *O26* allele was heterozygous in two samples. Other rare *O* alleles were found in heterozygous samples: *O28*, which is similar to *O01* except for a rare (926A>G, exon 7) SNP (Roubinet et al., 2001) with frequency of a little over 0% in Europe (according to NCBI); *O63*; *O73*, which differs from *O02* by 1 missense SNP (595 C>T/Arg199Cys) in exon 7 with 220C (not carrying the *O02* SNP 220C>T), and *O75* allele. Interestingly, a suggested novel *O* allele was found in one sample with a heterozygous missense SNP 482G>A (Arg161His) in exon 7. Although this SNP has been listed and validated in NCBI dbSNP, with frequency of 0% in Europe, it has not been allocated to an *ABO* allele as it is not listed in the dbRBC database (BGMUT) of 381 alleles (accessed July 2016).

In all the 13 B phenotype samples, *B101* was found heterozygous, which differs from the *A* allele by seven SNPs: in exon 6, 526C>G (Arg176Gly); in exon 7 703G>A (Gly235Ser), 796C>A (Leu266Met) and 803G>C (Gly268Ala). The SNPs 297A>G (exon 6), 657C>T and 930G>A (exon 7) are silent mutations, while the others are missense mutations. Of the 13 B samples, 9 were heterozygous for the *O01* allele, 2/13 were heterozygous with *O02* and the rest carried *O63* and *O75*.

Regarding samples of AB phenotype (8/47), the genotypes for 7 were *A201/B101* and one was *A101/B101*, leading to the 7 samples of A_2B phenotype and 1 of A_1B phenotype.

These mutations observed from the NGS genotyping required no further confirmation as data showed high coverage depth (650X) and quality; moreover, all identified mutations were reported and validated in the NCBI (dbSNP) with corresponding alleles listed in dbRBC databases (BGMUT), apart from the novel *O* allele found here.

Table 5.6 NGS genotyping of 47 samples of different ABO phenotypes. *(Note: samples of the same phenotype were grouped in tables in the next pages).*

The 47 samples (all with provided serological phenotypes) were sequenced by *ABO*-specific LR-PCR. The samples were ordered according to the phenotype: as A, O, B and AB. The genotyping results, shown for all 7 exons of the *ABO* gene, revealed several rare *O* alleles, such as *O26*, *O28*, *O63*, *O73* and *O75*, in addition to a possible novel *O* allele (found in an O phenotype sample) with a missense SNP 482G>A (Arg161His). The underlined mutations relate to those underlined alleles. The NHSBT provided and predicted phenotypes according to the genotyping are shown, which matched apart from the more detailed phenotype of A_2 and A_2B, which were revealed by NGS genotyping. The allele names were in accordance with dbRBC and ISBT (in brackets). The genotyping data was assessed as against the consensus reference sequence (*A101*).

Sample No.	Provided Phenotype/ Phenotype *	Possible Genotype	Exon 2	Exon 3	Exon 4	Exon 5	Exon 6	Exon 7
01,02	**A/A_1**	***A101(ABO*A1.01)/O26***					**88fs 261delG (Het)**	**Ile256_768C>A (Het)**
03,04	**A/ A_1**	***A101(ABO*A1.01)/ O02(ABO*O.01.02)***		Val 36Phe 106G>T (Het)	Arg63 189C>T (Het), Arg63His 188G>A (Het),	**Pro74Ser 220C>T (Het)**	**Thr99 297A>G (Het), 88fs 261delG (Het)**	Val277Met/829G>A (Het), Pro257 771C>T (Het), Pro227 681G>A (Het), Phe216Ile 646T>A (Het)
05	**A/ A_1**	***A101(ABO*A1.01)/ O01(ABO*O.01.01)***					**261delG (Het)**	
06	**A/ A_1**	***A102(ABO*A1.02)/ O02(ABO*O.01.02)***		Val 36Phe 106G>T (Het)	Arg63 189C>T (Het), Arg63His 188G>A (Het)	**Pro74Ser 220C>T (Het)**	**Thr99 297A>G (Het), 88fs 261delG (Het)**	Val277Met/829G>A (Het), Pro257 771C>T (Het), Pro227 681G>A (Het), Phe216Ile 646T>A (Het), **Pro156Leu 467C>T (Het)**
07	**A/ A_1**	***A101(ABO*A1.01)/O63***		Val 36Phe 106G>T (Het), **Gly35Arg 103G>A (Het)**	Arg63 189C>T (Hom), Arg63His 188G>A (Het)	**Pro74Ser 220C>T (Het)**	**Thr99 297A>G (Het), 88fs 261delG (Het)**	Val277Met/829G>A (Het), Pro257 771C>T (Het), Pro227 681G>A (Het), Phe216Ile 646T>A (Het)
08	**A/A_1**	***A101(ABO*A1.01)/ A201(ABO*A2.01)***						**1061delC (Het), Pro156Leu 467C>T (Het)**
09	**A/A_2**	***A201(ABO*A2.01)/O75***		**Val 36Phe** 106G>T (Het)	Arg63 **189C>T** (Het), **Arg63His** 188G>A (Het),	**Pro74Ser** 220C>T (Het)	Thr99 **297A>G (Het), 88fs 261delG (Het)**	**1061delC (Het),** **Val277Met**/829G>A (Het), Pro257 **771C>T** (Het), Pro227 **681G>A** (Het), **Phe216Ile 646T>A** (Het), **Trp181stop 542G>A (Het), Pro156Leu** 467C>T (Het)
10	**A/A_2**	***A201(ABO*A2.01)/ A201(ABO*A2.01)***						**1061delC (Hom), Pro156Leu 467C>T (Hom)**
11	**A/ A_2**	***A201(ABO*A2.01)/ O03(ABO*O.02.01)***	Arg18Leu 53G>T (Het)			**Pro74Ser 220C>T (Het)**	**Thr99 297A>G (Het)**	**1061delC (Het), Pro156Leu 467C>T (Het), Gly268Arg 802G>A (Het), Arg176Gly 526C>G (Het)**

Sample No.	Provided Phenotype/ Phenotype *	Possible Genotype	Exon 2	Exon 3	Exon 4	Exon 5	Exon 6	Exon 7
12	**O**	***O63/O (NOVEL)***		Val 36Phe 106G>T (Het), **Gly35Arg 103G>A (Het)**	Arg63 189C>T (Het), Arg63His 188G>A (Het),	**Pro74Ser 220C>T (Het)**	**Thr99 297A>G (Het), 88fs 261delG (Hom)**	Val277Met/829G>A (Het), Pro257 771C>T (Het), Pro227 681G>A (Het), Phe216Ile 646T>A (Het), **Arg161His 482G>A (Het)**
13,16, 19, 22, 24, 25	**O**	***O01(ABO*O.01.01)/ O01(ABO*O.01.01)***					**88fs 261delG (Hom)**	
14, 20	**O**	***O01(ABO*O.01.01)/ O75***		Val 36Phe 106G>T (Het)	Arg63 189C>T (Het), Arg63His 188G>A (Het),	**Pro74Ser 220C>T (Het)**	**Thr99 297A>G (Het), 88fs 261delG (Hom)**	Val277Met/829G>A (Het), Pro257 771C>T (Het), Pro227 681G>A (Het), Phe216Ile 646T>A (Het), **Trp181stop 542G>A (Het)**
15	**O**	***O02(ABO*O.01.02)/ O73***		Val 36Phe 106G>T (Hom)	Arg63 189C>T (Hom), Arg63His 188G>A (Hom),	**Pro74Ser 220C>T (Het)**	**Thr99 297A>G (Hom), 88fs 261delG (Hom)**	Val277Met/829G>A (Hom), Pro257 771C>T (Hom), Pro227 681G>A (Hom), Phe216Ile 646T>A (Hom), **Arg199Cys 595C>T (Het)**
17	**O**	***O26/ O68(ABO*O.02.17.1)***		Val 36Phe 106G>T (Het)	Arg63 189C>T (Het), Arg63His 188G>A (Het),		**Thr99 297A>G (Het), 88fs 261delG (Hom)**	Val277Met/829G>A (Het), Pro257 771C>T (Het), **Ile256 768C>A (Het),** Pro227 681G>A (Het), Phe216Ile 646T>A (Het)
18	**O**	***O03(ABO*O.02.01)/ O28***	Arg18Leu 53G>T (Het)			**Pro74Ser 220C>T (Het)**	**Thr99 297A>G (Het), 88fs 261delG (Het)**	**Tyr309Cys 926A>G (Het) Gly268Arg 802G>A** (Het), **Arg176Gly 526C>G (Het)**
21	**O**	***O01(ABO*O.01.01)/ O03(ABO*O.02.01)***	Arg18Leu 53G>T (Het)			**Pro74Ser 220C>T (Het)**	**Thr99 297A>G (Het), 88fs 261delG (Het)**	**Gly268Arg 802G>A (Het) Arg176Gly 526C>G (Het)**
23	**O**	***O01(ABO*O.01.01)/ O26***					**88fs 261delG (Hom)**	**Ile256 768C>A (Het)**
26	**O**	***O02(ABO*O.01.02)/ O02(ABO*O.01.02)***		Val 36Phe 106G>T (Hom)	Arg63 189C>T (Hom), Arg63His 188G>A (Hom)	**Pro74Ser 220C>T (Hom)**	**Thr99 297A>G (Hom), 88fs 261delG (Hom)**	Val277Met/829G>A (Hom), Pro257 771C>T (Hom), Pro227 681G>A (Hom), Phe216Ile 646T>A (Hom),

Sample No.	Provided Phenotype/Phenotype*	Possible Genotype	Exon 2	Exon 3	Exon 4	Exon 5	Exon 6	Exon 7
27, 28, 29, 30, 31, 32, 34, 36, 38	**B**	***B101(ABO*B.01)/ O01(ABO*O.01.01)***					**Thr99 297A>G (Het), 88fs 261delG (Het)**	Leu310 **930G>A** (Het), **Gly268Ala** 803G>C (Het), **Leu266Met** 796C>A (Het), **Gly235Ser** 703G>A (Het), His219 **657C>T** (Het), **Arg176Gly** 526C>G (Het)
33	**B**	***B101(ABO*B.01)/ O63***		Val 36Phe 106G>T (Het), **Gly35Arg 103G>A (Het)**	Arg63 189C>T (Het), Arg63His 188G>A (Het),	**Pro74Ser 220C>T (Het)**	**Thr99 297A>G (Hom), 88fs 261delG (Het)**	Leu310 930G>A (Het), Val277Met/829G>A (Het), Gly268Ala 803G>C (Het), Leu266Met 796C>A (Het), Pro257 771C>T (Het), Gly235Ser 703G>A (Het), Pro227 681G>A (Het), His219 657C>T (Het), Phe216Ile 646T>A (Het), Arg176Gly 526C>G (Het)
35	**B**	***B101(ABO*B.01)/ O75***		Val 36Phe 106G>T (Het)	Arg63 189C>T (Het), Arg63His 188G>A (Het),	**Pro74Ser 220C>T (Het)**	**Thr99 297A>G (Hom), 88fs 261delG (Het)**	Leu310 930G>A (Het), Val277Met/829G>A (Het), Gly268Ala 803G>C (Het), Leu266Met 796C>A (Het), Pro257 771C>T (Het), Gly235Ser 703G>A (Het), Pro227 681G>A (Het), His219 657C>T (Het), Phe216Ile 646T>A (Het), **Trp181stop 542G>A (Het),** Arg176Gly 526C>G (Het)
37, 39	**B**	***B101(ABO*B.01)/ O02(ABO*O.01.02)***		Val 36Phe 106G>T (Het)	Arg63 189C>T (Het), Arg63His 188G>A (Het),	**Pro74Ser 220C>T (Het)**	**Thr99 297A>G (Hom), 88fs 261delG (Het)**	Leu310 930G>A (Het), Val277Met/829G>A (Het), Gly268Ala 803G>C (Het), Leu266Met 796C>A (Het), Pro257 771C>T (Het), Gly235Ser 703G>A (Het), Pro227 681G>A (Het), His219 657C>T (Het), Phe216Ile 646T>A (Het), Arg176Gly 526C>G (Het)

Sample No.	Provided Phenotype/Phenotype*	Possible Genotype	Exon 2	Exon 3	Exon 4	Exon 5	Exon 6	Exon 7
40, 41, 42, 43, 44, 45, 46	**AB/A_2B**	***A201(ABO*A2.01)/ B101(ABO*B.01)***					**Thr99 297A>G (Het)**	**1061delC (Het),** Leu310 930G>A (Het), Gly268Ala 803G>C (Het), Leu266Met 796C>A (Het), Gly235Ser 703G>A (Het), His219 657C>T (Het), Arg176Gly 526C>G (Het), **Pro156Leu 467C>T (Het)**
47	**AB/A_1B**	***A101(ABO*A1.01)/ B101(ABO*B.01)***					**Thr99 297A>G (Het)**	Leu310 930G>A (Het), Gly268Ala 803G>C (Het), Leu266Met 796C>A (Het), Gly235Ser 703G>A (Het), His219 657C>T (Het), Arg176Gly 526C>G (Het),

5.3.6.2 *ABO* mutations in introns

NGS genotyping revealed all existing *ABO* gene mutations, including those in introns and flanking regions. This may be useful in studies of allele evolution and forensics for better resolution of the molecular basis of the *ABO* allele. A significant number of intronic SNPs and indels were found, with an average of 74 mutations in samples of A phenotype, 59 in O, 79 in B and 96 in AB. Correlation of these intronic mutations with the allele-defining exonic SNPs in our samples was challenging, especially when this task was carried out mostly manually. The reason for this was the small number of samples carrying homozygous alleles: 1 (*A201*/*A201*), 6 *O01* and 1 *O02*. Table 5.7 lists the intronic mutations analysed across the *ABO* gene. The analysis of the intronic mutations was (green for homozygous match the reference, orange homozygous different and crossed blue heterozygous). The reference sequence is suggested to represent both *O01* and *O02*, which was seen in intronic and exonic mutations. Aligning the samples with homozygous *O* alleles (*O01* or *O02*) (Table 5.8) showed this phenomenon, as in intron 4 there were insertions of a C nucleotide and an extra 13bp at chromosomal positions chr9:136133583 and 136135138-chr9:136135150, respectively. While this region of the reference sequence was in concordance with the *O02* allele, the sample of *O02/O02* genotype showed no difference. In contrast, *O01/O01* samples showed a homozygous deletion which was not shown in the reference sequence. There was also a CCC insertion at chr9:136133381-83 in intron 5 along with intronic SNPs within intron 2-5 and intron 6-7 (Table 5.7 and 5.8). From Table 5.8, it can be seen that there is correlation of intronic SNPs across the gene, especially in introns 3 to 6, with homozygous samples containing *O01* and *O02* alleles, while sample ABO001.14 (*O01/O75*) represented the heterozygous sample. On the other hand, an 8 nucleotide deletion was detected in chr9:136139908-136139915 in intron 1 in all samples with *A* alleles (*A1* and *A2*) (11 A phenotype samples and 8 AB phenotype samples). Only

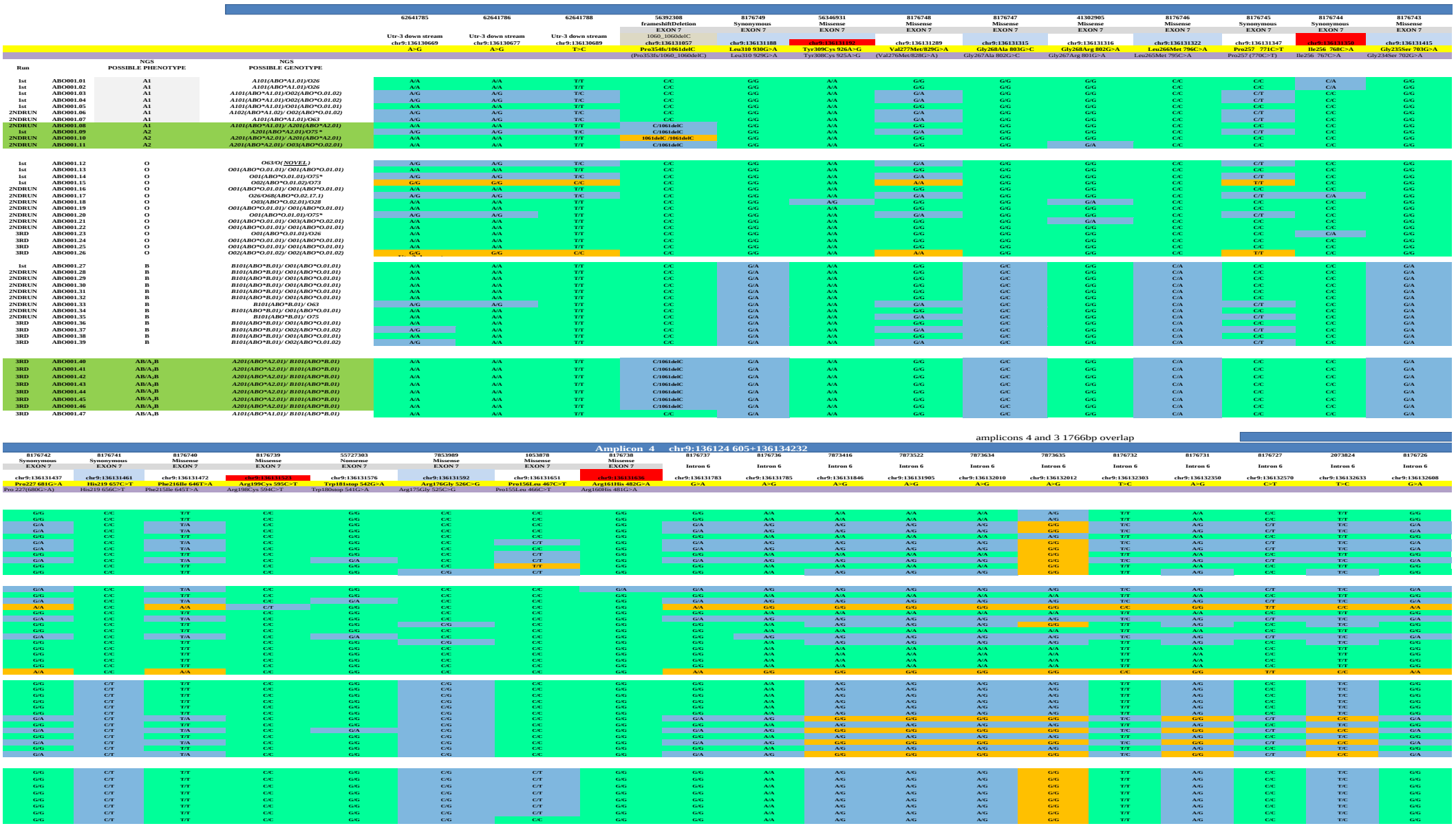
samples carrying the *A* allele were different to the reference sequence; samples ABO001.8 and .10 were homozygously different as the genotype were (*A101*/*A201* and *A201*/*A201*, respectively) (Table 5.7). With regards to *O* alleles, the pattern of the intronic SNPs show similarity of *O26* and *O28* alleles heterozygous with *O01*, while *O63*, *-68*, *-73*, and -75 were similar to *O02* (Table 5.7), which illustrates the benefit of this approach in analysing the evolution of *ABO* alleles. A 22 bp deletion was found in samples carrying the *A201* allele (homozygous in samples ABO001.08, .10 and .11) and heterozygous in sample ABO001.09). All samples the with AB phenotype (7 A_2B and 1 A_1B) were homozygous for this deletion, which also applies to 2 intronic SNPs (in chr9: 136145118 and 13614597 adjacent to this deletion). Nevertheless, this deletion and the intronic SNPs were absent in other samples of A, B and O phenotype, which might suggest the possibility of an allelic dropout in amplicon 2 during the PCR amplification. However, analysis of all 47 samples for polymorphisms around the primer pair sites for amplicon 2 showed no visible SNPs, which suggested that the amplification conditions might need optimisation.

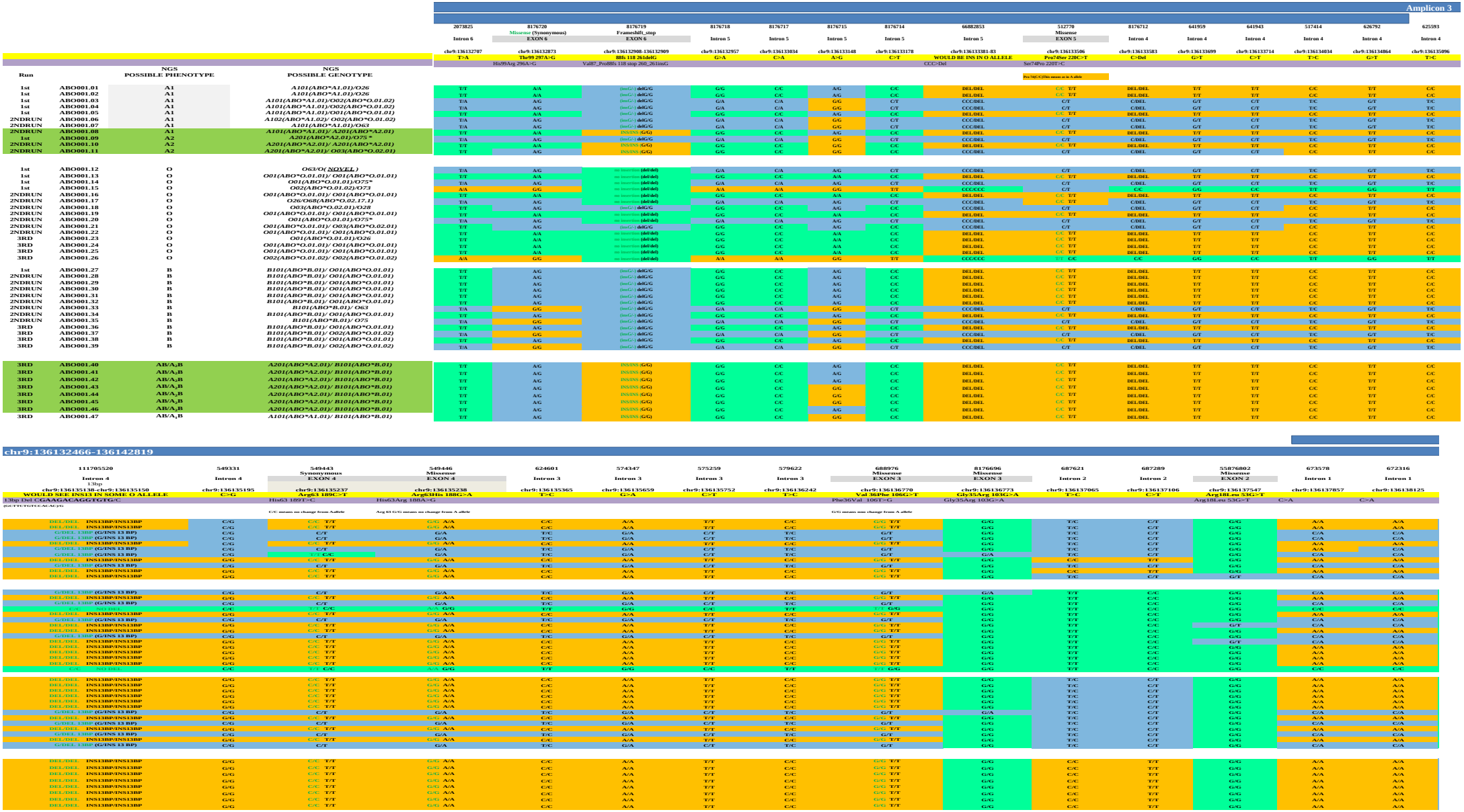
The upstream area that includes the CBF/NF-Y transcription factor binding site with four (43 bp) repeats about 3.8kb upstream from the start codon was analysed. The 43 bp sequence is ACCCCAGCCAATAGGGGAAGGACACAGAAACAGAAACTGCGTT (Kominato et al., 1997, Irshaid et al., 1999). The four repeats of the 43 bp sequence were all covered by the NGS sequence data for all samples irrespective of phenotype. There were 2 SNPs in the nt. 41 in 2 repeats were inconsistent between different *ABO* alleles. The first repeat appeared to carry C in nucleotide 41 in the reference sequence (chromosomal location ch9: 136154433), which suggested to be *O02*. This is seen in sample ABO001.26 with genotype *O02(ABO*O.01.02)/O02(ABO*O.01.02)* with no change from the reference allele (Table 5.9). However, there was no consistency with other alleles, *A, B* and *O01* (Table 5.9). Another SNP (G>A), which has been described

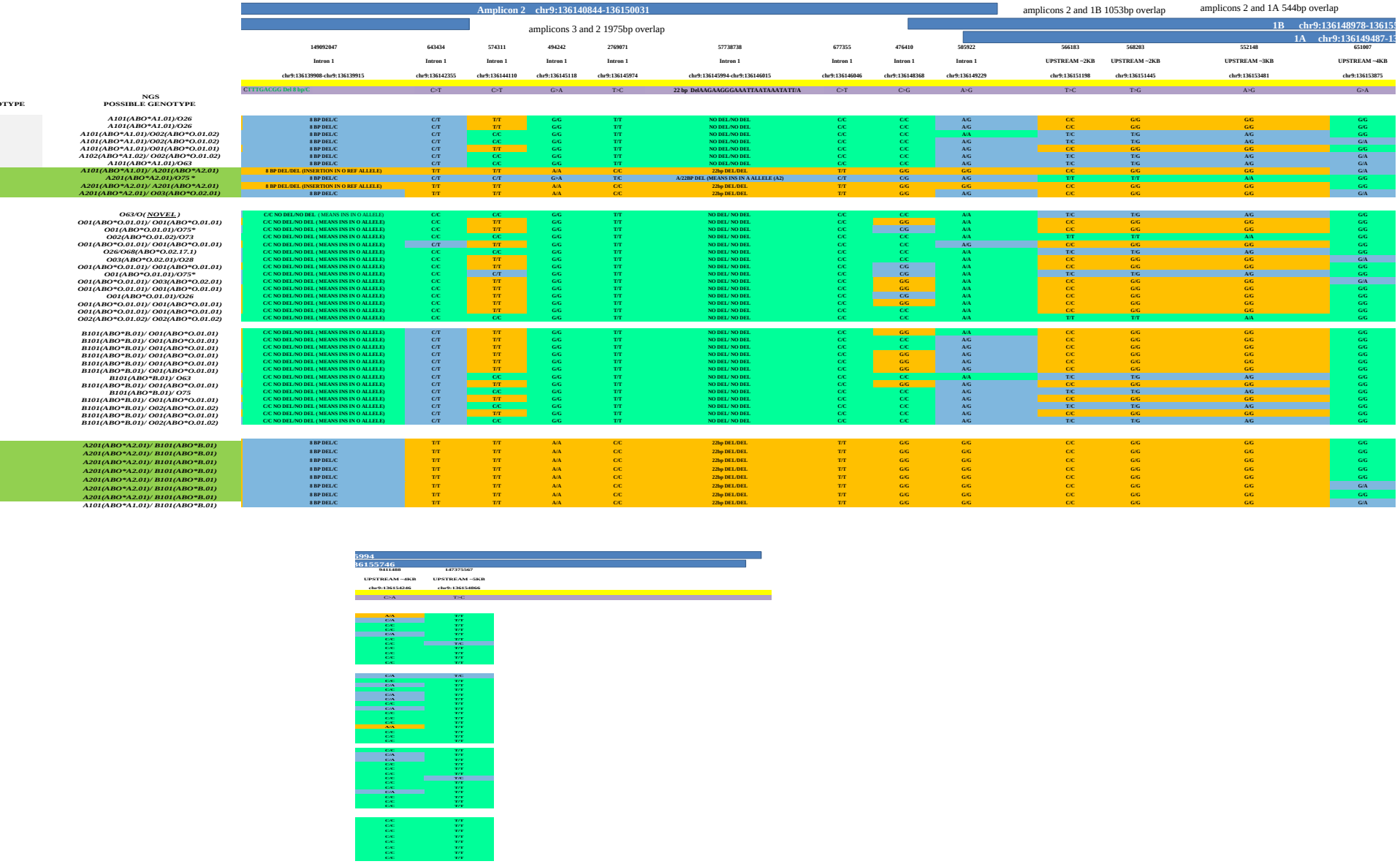
before (Irshaid et al., 1999) to be carried in *A101/A101* or *A101/O02* genotype samples, was found in 6/11 samples of A phenotype but not with complete consistency as it was seen in samples ABO001.03 and .04 samples: these samples are of the same suggested genotype (*A101(ABO*A1.01)/O02(ABO*O.01.02*), but only the latter was heterozygous with the SNP. In addition, this SNP was not found in sample ABO001.26, the genotype of which (*O02/O02*) may contradict the association of this SNP with the *O02* allele, as previously described (Irshaid et al., 1999). This SNP was not found in samples of B phenotype, but was heterozygous in 2/8 AB samples (A_2B and A_1B). This approach may thus reveal more specific intronic polymorphisms if applied in more samples with homozygous alleles.

Table 5.7 NGS genotyping of 47 ABO samples (all provided with serological phenotype) (Note: the table is distributed on the next pages due to its size).

A graphical illustration of the various mutations found in the ABO samples from NGS genotyping. The green represents homozygous resemblance to the reference; orange represents homozygous difference to the reference; while the crossed blue represents heterozygosity. Polymorphisms in exons are coloured grey, while red represents the rare mutation listed in the dbRBC and (NCBI/1000 genome), the rsID for which is shown above the polymorphisms. The yellow highlighted line is the consensus annotation. The four overlapping amplicons are shown as blue bars on the top. Due to the lack of the number of homozygous samples, the correlation study of these SNPs with the alleles was challenging. Larger tables in appendix B.


amplicons 4 and 3 1766bp overlap
Amplicon 4 chr9:136124 605+136134232


Amplicon 3
chr9:136132707 T>A Thr99 297A>G
chr9:136132873 His99Arg 296A>G
chr9:136132908-136132909 88ls 118 261delG Val87_Pro88ls 118 stop 260_261insG
chr9:136132957 G>A
chr9:136133034 C>A
chr9:136133148 A>G
chr9:136133178 C>T
chr9:136133381-83 WOULD BE INS IN O ALLELE
chr9:136133506 Pro745ar 228C>T Ser749Pro 220T>C
chr9:136133583 C>Del
chr9:136133699 G>T
chr9:136133714 C>T
chr9:1361334034 T>C
chr9:136134864 G>T
chr9:136135096 T>C
NGS POSSIBLE PHENOTYPE
NGS POSSIBLE GENOTYPE
chr9:136135138-chr9:136135150 WOULD SEE INS13 IN SOME O ALLELE
chr9:136135195 C>G
chr9:136135237 Arg63 189C>T
chr9:136135238 Arg63His 188G>A
chr9:136135365 T>C
chr9:136135659 G>A
chr9:136135752 C>T
chr9:136136242 T>C
chr9:136136770 Val 36Phe 106G>T
chr9:136136773 Gly35Arg 103G>A
chr9:136137065 T>C
chr9:136137106 C>T
chr9:136137547 Arg18Leu 53G>T
chr9:136137857 C>A
chr9:136138125 C>A
chr9:136132466-136142819



Run		NGS POSSIBLE PHENOTYPE	NGS POSSIBLE GENOTYPE
1st	ABO001.01	A1	*A101(ABO*A1.01)/O26*
1st	ABO001.02	A1	*A101(ABO*A1.01)/O26*
1st	ABO001.03	A1	*A101(ABO*A1.01)/O02(ABO*O.01.02)*
1st	ABO001.04	A1	*A101(ABO*A1.01)/O02(ABO*O.01.02)*
1st	ABO001.05	A1	*A101(ABO*A1.01)/O01(ABO*O.01.01)*
2NDRUN	ABO001.06	A1	*A102(ABO*A1.02)/ O02(ABO*O.01.02)*
2NDRUN	ABO001.07	A1	*A101(ABO*A1.01)/O63*
2NDRUN	ABO001.08	A1	*A101(ABO*A1.01)/ A201(ABO*A2.01)*
1st	ABO001.09	A2	*A201(ABO*A2.01)/O75 **
2NDRUN	ABO001.10	A2	*A201(ABO*A2.01)/ A201(ABO*A2.01)*
2NDRUN	ABO001.11	A2	*A201(ABO*A2.01)/ O03(ABO*O.02.01)*
1st	ABO001.12	O	*O63/O(**NOVEL**)*
1st	ABO001.13	O	*O01(ABO*O.01.01)/ O01(ABO*O.01.01)*
1st	ABO001.14	O	*O01(ABO*O.01.01)/O75**
1st	ABO001.15	O	*O02(ABO*O.01.02)/O73*
2NDRUN	ABO001.16	O	*O01(ABO*O.01.01)/ O01(ABO*O.01.01)*
2NDRUN	ABO001.17	O	*O26/O68(ABO*O.02.17.1)*
2NDRUN	ABO001.18	O	*O03(ABO*O.02.01)/O28*
2NDRUN	ABO001.19	O	*O01(ABO*O.01.01)/ O01(ABO*O.01.01)*
2NDRUN	ABO001.20	O	*O01(ABO*O.01.01)/O75**
2NDRUN	ABO001.21	O	*O01(ABO*O.01.01)/ O03(ABO*O.02.01)*
2NDRUN	ABO001.22	O	*O01(ABO*O.01.01)/ O01(ABO*O.01.01)*
3RD	ABO001.23	O	*O01(ABO*O.01.01)/O26*
3RD	ABO001.24	O	*O01(ABO*O.01.01)/ O01(ABO*O.01.01)*
3RD	ABO001.25	O	*O01(ABO*O.01.01)/ O01(ABO*O.01.01)*
3RD	ABO001.26	O	*O02(ABO*O.01.02)/ O02(ABO*O.01.02)*
1st	ABO001.27	B	*B101(ABO*B.01)/ O01(ABO*O.01.01)*
2NDRUN	ABO001.28	B	*B101(ABO*B.01)/ O01(ABO*O.01.01)*
2NDRUN	ABO001.29	B	*B101(ABO*B.01)/ O01(ABO*O.01.01)*
2NDRUN	ABO001.30	B	*B101(ABO*B.01)/ O01(ABO*O.01.01)*
2NDRUN	ABO001.31	B	*B101(ABO*B.01)/ O01(ABO*O.01.01)*
2NDRUN	ABO001.32	B	*B101(ABO*B.01)/ O01(ABO*O.01.01)*
2NDRUN	ABO001.33	B	*B101(ABO*B.01)/ O63*
2NDRUN	ABO001.34	B	*B101(ABO*B.01)/ O01(ABO*O.01.01)*
2NDRUN	ABO001.35	B	*B101(ABO*B.01)/ O75*
3RD	ABO001.36	B	*B101(ABO*B.01)/ O01(ABO*O.01.01)*
3RD	ABO001.37	B	*B101(ABO*B.01)/ O02(ABO*O.01.02)*
3RD	ABO001.38	B	*B101(ABO*B.01)/ O01(ABO*O.01.01)*
3RD	ABO001.39	B	*B101(ABO*B.01)/ O02(ABO*O.01.02)*
3RD	ABO001.40	AB/A_2B	*A201(ABO*A2.01)/ B101(ABO*B.01)*
3RD	ABO001.41	AB/A_2B	*A201(ABO*A2.01)/ B101(ABO*B.01)*
3RD	ABO001.42	AB/A_2B	*A201(ABO*A2.01)/ B101(ABO*B.01)*
3RD	ABO001.43	AB/A_2B	*A201(ABO*A2.01)/ B101(ABO*B.01)*
3RD	ABO001.44	AB/A_2B	*A201(ABO*A2.01)/ B101(ABO*B.01)*
3RD	ABO001.45	AB/A_2B	*A201(ABO*A2.01)/ B101(ABO*B.01)*
3RD	ABO001.46	AB/A_2B	*A201(ABO*A2.01)/ B101(ABO*B.01)*
3RD	ABO001.47	AB/A_2B	*A101(ABO*A1.01)/ B101(ABO*B.01)*

Table 5.8 NGS genotyping of homozygous *O* allele samples (Note: the table is on the next pages due to its size).

A graphical illustration of the correlation of intronic polymorphisms with homozygous *O* alleles obtained from the analysis of the NGS data. The polymorphisms of 6 *O01* homozygous and *1 O01/O26* alleles (of which the latter resembles *O01*) are shown to be correlated. A *O02* homozygous sample was also displayed. One sample of genotype *O01/O75*, the latter allele resembles *O02*, was compared to the homozygous samples, which shows the heterozygosity of the polymorphisms. The correlation was mainly in the intronic polymorphisms in intron 3-6. The green represents homozygous resemblance to the reference; orange represents homozygous difference to the reference; while the crossed blue represents heterozygosity. Polymorphisms in exons are coloured grey, while red represents the rare mutation listed in the dbRBC and (NCBI/1000 genome), the rsID for which is showed above the polymorphisms. The yellow highlighted line is the consensus annotation. Larger tables in appendix B.

			62641785	62641786	62641788	56392308	8176749	56346931	8176748	8176747	41302905	8176746	8176745	8176744	8176743	8176742	8176741
						frameshiftDeletion	Synonymous	Missense	Missense	Missense	Missense	Missense	Synonymous	Synonymous	Missense	Synonymous	Synonymous
						EXON 7	EXON 7	EXON 7	EXON 7	EXON 7	EXON 7	EXON 7	EXON 7	EXON 7	EXON 7	EXON 7	EXON 7
			Utr-3 down stream	Utr-3 down stream	Utr-3 down stream	1060_1060delC											
			chr9:136130669	chr9:136130677	chr9:136130689	chr9:136131057	chr9:136131188	chr9:136131192	chr9:136131289	chr9:136131315	chr9:136131316	chr9:136131322	chr9:136131347	chr9:136131350	chr9:136131415	chr9:136131437	chr9:136131461
			A>G	A>G	T>C	Pro354fs/1061delC	Leu310 930G>A	Tyr309Cys 926A>G	Val277Met/829G>A	Gly268Ala 803G>C	Gly268Arg 802G>A	Leu266Met 796C>A	Pro257 771C>T	Ile256 768C>A	Gly235Ser 703G>A	Pro227 681G>A	His219 657C>T
						(Pro353fs/1060_1060delC)	Leu310 929G>A	Tyr308Cys 925A>G	(Val276Met/828G>A)	Gly267Ala 802G>C	Gly267Arg 801G>A	Leu265Met 795C>A	Pro257 (770C>T)	Ile256 767C>A	Gly234Ser 702G>A	Pro 227(680G>A)	His219 656C>T
ABO001.23	0	*O01(ABO*O.01.01)/O26*	A/A	A/A	T/T	C/C	G/G	A/A	G/G	G/G	G/G	C/C	C/C	C/A	G/G	G/G	C/C
ABO001.13	0	*O01(ABO*O.01.01)/ O01(ABO*O.01.01)*	A/A	A/A	T/T	C/C	G/G	A/A	G/G	G/G	G/G	C/C	C/C	C/C	G/G	G/G	C/C
ABO001.16	0	*O01(ABO*O.01.01)/ O01(ABO*O.01.01)*	A/A	A/A	T/T	C/C	G/G	A/A	G/G	G/G	G/G	C/C	C/C	C/C	G/G	G/G	C/C
ABO001.19	0	*O01(ABO*O.01.01)/ O01(ABO*O.01.01)*	A/A	A/A	T/T	C/C	G/G	A/A	G/G	G/G	G/G	C/C	C/C	C/C	G/G	G/G	C/C
ABO001.22	0	*O01(ABO*O.01.01)/ O01(ABO*O.01.01)*	A/A	A/A	T/T	C/C	G/G	A/A	G/G	G/G	G/G	C/C	C/C	C/C	G/G	G/G	C/C
ABO001.24	0	*O01(ABO*O.01.01)/ O01(ABO*O.01.01)*	A/A	A/A	T/T	C/C	G/G	A/A	G/G	G/G	G/G	C/C	C/C	C/C	G/G	G/G	C/C
ABO001.25	0	*O01(ABO*O.01.01)/ O01(ABO*O.01.01)*	A/A	A/A	T/T	C/C	G/G	A/A	G/G	G/G	G/G	C/C	C/C	C/C	G/G	G/G	C/C
ABO001.26	0	*O02(ABO*O.01.02)/ O02(ABO*O.01.02)*	G/G	G/G	C/C	C/C	G/G	A/A	A/A	G/G	G/G	C/C	T/T	C/C	G/G	A/A	C/C
ABO001.14	0	*O01(ABO*O.01.01)/O75**	A/G	A/G	T/C	C/C	G/G	A/A	G/A	G/G	G/G	C/C	C/T	C/C	G/G	G/A	C/C

	8176740	8176739	55727303	7853989	1053878	8176738	8176737	8176736	7873416	7873522	7873634	7873635	8176732	8176731	8176727	2073824	8176726	2073825	8176720
	Missense	Missense	Nonsense	Missense	Missense	Missense													Missense (Synonymous)
	EXON 7	EXON 7	EXON 7	EXON 7	EXON 7	EXON 7	Intron 6	Intron 6	Intron 6	Intron 6	Intron 6	Intron 6	Intron 6	Intron 6	Intron 6	Intron 6	Intron 6	Intron 6	EXON 6
	chr9:136131472	chr9:136131523	chr9:136131576	chr9:136131592	chr9:136131651	chr9:136131636	chr9:136131783	chr9:136131785	chr9:136131846	chr9:136131905	chr9:136132010	chr9:136132012	chr9:136132303	chr9:136132350	chr9:136132570	chr9:136132633	chr9:136132608	chr9:136132707	chr9:136132873
	Phe216Ile 646T>A	Arg199Cys 595C>T	Trp181stop 542G>A	Arg176Gly 526C>G	Pro156Leu 467C>T	Arg161His 482G>A	G>A	A>G	A>G	A>G	A>G	A>G	T>C	A>G	C>T	T>C	G>A	T>A	Thr99 297A>G
	Phe215Ile 645T>A	Arg198Cys 594C>T	Trp180stop 541G>A	Arg175Gly 525C>G	Pro155Leu 466C>T	Arg160His 481G>A													His99Arg 296A>G
	T/T	C/C	G/G	C/C	C/C	G/G	G/G	A/A	A/A	A/A	A/A	A/A	T/T	A/A	C/C	T/T	G/G	T/T	A/A
	T/T	C/C	G/G	C/C	C/C	G/G	G/G	A/A	A/A	A/A	A/A	A/A	T/T	A/A	C/C	T/T	G/G	T/T	A/A
	T/T	C/C	G/G	C/C	C/C	G/G	G/G	A/A	A/A	A/A	A/A	A/A	T/T	A/A	C/C	T/T	G/G	T/T	A/A
	T/T	C/C	G/G	C/C	C/C	G/G	G/G	A/A	A/A	A/A	A/A	A/A	T/T	A/A	C/C	T/T	G/G	T/T	A/A
	T/T	C/C	G/G	C/C	C/C	G/G	G/G	A/A	A/A	A/A	A/A	A/A	T/T	A/A	C/C	T/T	G/G	T/T	A/A
	T/T	C/C	G/G	C/C	C/C	G/G	G/G	A/A	A/A	A/A	A/A	A/A	T/T	A/A	C/C	T/T	G/G	T/T	A/A
	T/T	C/C	G/G	C/C	C/C	G/G	G/G	A/A	A/A	A/A	A/A	A/A	T/T	A/A	C/C	T/T	G/G	T/T	A/A
	A/A	C/C	G/G	C/C	C/C	G/G	A/A	G/G	G/G	G/G	G/G	G/G	C/C	G/G	T/T	C/C	A/A	A/A	G/G
	T/A	C/C	G/A	C/C	C/C	G/G	G/A	A/G	A/G	A/G	A/G	A/G	T/C	A/G	C/T	T/C	G/A	T/A	A/G

	8176719	8176718	8176717	8176715	8176714	66882853	512770	8176712	641959	641943	517414	626792	625593	111705520	549331	549443	549446
	Frameshift_stop						Missense									Synonymous	Missense
	EXON 6	Intron 5	Intron 5	Intron 5	Intron 5	Intron 5	EXON 5	Intron 4	Intron 4	Intron 4	Intron 4	Intron 4	Intron 4	Intron 4	Intron 4	EXON 4	EXON 4
	chr9:136132908-136132909	chr9:136132957	chr9:136133034	chr9:136133148	chr9:136133178	chr9:136133381-83	chr9:136133506	chr9:136133583	chr9:136133699	chr9:136133714	chr9:136134034	chr9:136134864	chr9:136135096	13bp chr9:136135138-chr9:136135150	chr9:136135195	chr9:136135237	chr9:136135238
	88fs 118 261delG	G>A	C>A	A>G	C>T	WOULD BE INS IN O ALLELE	Pro74Ser 220C>T	C>Del	G>T	C>T	T>C	G>T	T>C	WOULD SEE INS13 IN SOME O ALLELE	C>G	Arg63 189C>T	Arg63His 188G>A
	Val87_Pro88fs 118 stop 260_261insG					CCC>Del	Ser74Pro 220T>C							13bp Del CGAAGACAGGTGTG/C		His63 189T>C	His63Arg 188A>G
	no insertion (del/del)	G/G	C/C	A/A	C/C	DEL/DEL	C/C T/T	DEL/DEL	T/T	T/T	C/C	T/T	C/C	DEL/DEL INS13BP/INS13BP	G/G	C/C T/T	G/G A/A
	no insertion (del/del)	G/G	C/C	A/A	C/C	DEL/DEL	C/C T/T	DEL/DEL	T/T	T/T	C/C	T/T	C/C	DEL/DEL INS13BP/INS13BP	G/G	C/C T/T	G/G A/A
	no insertion (del/del)	G/G	C/C	A/A	C/C	DEL/DEL	C/C T/T	DEL/DEL	T/T	T/T	C/C	T/T	C/C	DEL/DEL INS13BP/INS13BP	G/G	C/C T/T	G/G A/A
	no insertion (del/del)	G/G	C/C	A/A	C/C	DEL/DEL	C/C T/T	DEL/DEL	T/T	T/T	C/C	T/T	C/C	DEL/DEL INS13BP/INS13BP	G/G	C/C T/T	G/G A/A
	no insertion (del/del)	G/G	C/C	A/A	C/C	DEL/DEL	C/C T/T	DEL/DEL	T/T	T/T	C/C	T/T	C/C	DEL/DEL INS13BP/INS13BP	G/G	C/C T/T	G/G A/A
	no insertion (del/del)	G/G	C/C	A/A	C/C	DEL/DEL	C/C T/T	DEL/DEL	T/T	T/T	C/C	T/T	C/C	DEL/DEL INS13BP/INS13BP	G/G	C/C T/T	G/G A/A
	no insertion (del/del)	A/A	A/A	G/G	T/T	CCC/CCC	T/T C/C	C/C	G/G	C/C	T/T	G/G	T/T	C/C NO DEL	C/C	T/T C/C	A/A G/G
	no insertion (del/del)	G/A	C/A	A/G	C/T	CCC/DEL	C/T	C/DEL	G/T	C/T	T/C	G/T	T/C	G/DEL 13BP (G/INS 13 BP)	C/G	C/T	G/A

			624601	574947	575259	579622	608976	8176696	687621	687289	55876802	673578	672316	149082047	643434	574311
							Missense	Missense			Missense					
			Intron 3	Intron 3	Intron 3	Intron 3	EXON 3	EXON 3	Intron 2	Intron 2	EXON 2	Intron 1	Intron 1	Intron 1	Intron 1	Intron 1
			chr9:136135365	chr9:136135659	chr9:136135752	chr9:136136242	chr9:136136770	chr9:136136773	chr9:136137065	chr9:136137106	chr9:136137547	chr9:136137857	chr9:136138125	chr9:136139908-chr9:136139915	chr9:136142355	chr9:136144110
			T>C	G>A	C>T	T>C	Val 36Phe 106G>T	Gly35Arg 103G>A	T>C	C>T	Arg18Leu 53G>T					
							Phe36Val 106T>G					C>A	C>A	CTTTGACGG.Del 8 bp/C	i>T	C>T
ABO001.23	**O**	***O01(ABO*O.01.01)/O26***	CC	AA	TT	CC	GG TT	GG	TT	CC	GG	AA	AA	CC NO DEL/NO DEL (MEANS INS IN O ALLELE)	CC	TT
ABO001.13	**O**	***O01(ABO*O.01.01)/ O01(ABO*O.01.01)***	CC	AA	TT	CC	GG TT	GG	TT	CC	GG	AA	AA	CC NO DEL/NO DEL (MEANS INS IN O ALLELE)	CC	TT
ABO001.16	**O**	***O01(ABO*O.01.01)/ O01(ABO*O.01.01)***	CC	AA	TT	CC	GG TT	GG	TT	CC	GG	AA	AA	CC NO DEL/NO DEL (MEANS INS IN O ALLELE)	CT	TT
ABO001.19	**O**	***O01(ABO*O.01.01)/ O01(ABO*O.01.01)***	CC	AA	TT	CC	GG TT	GG	TT	CC	GG	AA	AA	CC NO DEL/NO DEL (MEANS INS IN O ALLELE)	CC	TT
ABO001.22	**O**	***O01(ABO*O.01.01)/ O01(ABO*O.01.01)***	CC	AA	TT	CC	GG TT	GG	TT	CC	GG	AA	AA	CC NO DEL/NO DEL (MEANS INS IN O ALLELE)	CC	TT
ABO001.24	**O**	***O01(ABO*O.01.01)/ O01(ABO*O.01.01)***	CC	AA	TT	CC	GG TT	GG	TT	CC	GG	AA	AA	CC NO DEL/NO DEL (MEANS INS IN O ALLELE)	CC	TT
ABO001.25	**O**	***O01(ABO*O.01.01)/ O01(ABO*O.01.01)***	CC	AA	TT	CC	GG TT	GG	TT	CC	GG	AA	AA	CC NO DEL/NO DEL (MEANS INS IN O ALLELE)	CC	TT
ABO001.26	**O**	***O02(ABO*O.01.02)/ O02(ABO*O.01.02)***	TT	GG	CC	TT	TT GG	GG	TT	CC	GG	CC	CC	CC NO DEL/NO DEL (MEANS INS IN O ALLELE)	CC	CC
ABO001.14	**O**	***O01(ABO*O.01.01)/O75****	TC	GA	CT	TC	G/T	GG	TT	CC	GG	CA	CA	CC NO DEL/NO DEL (MEANS INS IN O ALLELE)	CC	TT

494242	**2769071**	**57738738**	**677355**	**476410**	**505922**
Intron 1	**Intron 1**	**Intron 1**	**Intron 1**	**Intron 1**	**Intron 1**
chr9:136145118	**chr9:136145974**	**chr9:136145994-chr9:136146015**	**chr9:136146046**	**chr9:136148368**	**chr9:136149229**
G>A	T>C	22 bp Del**AAGAAGGGAAATTAATAAATATT**/A		C>G	A>G
G/G	**T/T**	**NO DEL/ NO DEL**	**C/C**	**C/G**	**A/A**
G/G	**T/T**	**NO DEL/ NO DEL**	**C/C**	**G/G**	**A/A**
G/G	**T/T**	**NO DEL/ NO DEL**	**C/C**	**C/C**	**A/G**
G/G	**T/T**	**NO DEL/ NO DEL**	**C/C**	**C/G**	**A/A**
G/G	**T/T**	**NO DEL/ NO DEL**	**C/C**	**G/G**	**A/A**
G/G	**T/T**	**NO DEL/ NO DEL**	**C/C**	**G/G**	**A/A**
G/G	**T/T**	**NO DEL/ NO DEL**	**C/C**	**C/C**	**A/A**
G/G	**T/T**	**NO DEL/ NO DEL**	**C/C**	**C/C**	**A/A**
G/G	**T/T**	**NO DEL/ NO DEL**	**C/C**	**C/G**	**A/A**

ID	Phenotype	*Genotype*	nt 41 1st repeat*	nt 41 4th repeat**
ABO001.01	**A1**	***A101(ABO*A1.01)/O26***	**C**	**G**
ABO001.02	**A1**	***A101(ABO*A1.01)/O26***	**C**	**G**
ABO001.03	**A1**	***A101(ABO*A1.01)/O02(ABO*O.01.02)***	**C**	**G**
ABO001.04	**A1**	***A101(ABO*A1.01)/O02(ABO*O.01.02)***	**C**	**G/A HET**
ABO001.05	**A1**	***A101(ABO*A1.01)/O01(ABO*O.01.01)***	**C**	**G/A HET**
ABO001.06	**A1**	***A102(ABO*A1.02)/ O02(ABO*O.01.02)***	**C**	**G/A HET**
ABO001.07	**A1**	***A101(ABO*A1.01)/O63***	**C**	**G/A HET**
ABO001.08	**A1**	***A101(ABO*A1.01)/ A201(ABO*A2.01)***	**G**	**G/A HET**
ABO001.09	**A2**	***A201(ABO*A2.01)/O75****	**C**	**G**
ABO001.10	**A2**	***A201(ABO*A2.01)/ A201(ABO*A2.01)***	**G**	**G**
ABO001.11	**A2**	***A201(ABO*A2.01)/ O03(ABO*O.02.01)***	**C/A Het**	**G/A HET**
ABO001.12	**O**	***O63/O(NOVEL)***	**C**	**G**
ABO001.13	**O**	***O01(ABO*O.01.01)/ O01(ABO*O.01.01)***	**G**	**G**
ABO001.14	**O**	***O01(ABO*O.01.01)/O75****	**C**	**G**
ABO001.15	**O**	***O02(ABO*O.01.02)/O73***	**C**	**G**
ABO001.16	**O**	***O01(ABO*O.01.01)/ O01(ABO*O.01.01)***	**G**	**G**
ABO001.17	**O**	***O26/O68(ABO*O.02.17.1)***	**C**	**G**
ABO001.18	**O**	***O03(ABO*O.02.01)/O28***	C/A HET	**G/A HET**
ABO001.19	**O**	***O01(ABO*O.01.01)/ O01(ABO*O.01.01)***	C/G HET	G
ABO001.20	**O**	***O01(ABO*O.01.01)/O75****	C/G HET	G
ABO001.21	**O**	***O01(ABO*O.01.01)/ O03(ABO*O.02.01)***	C/A HET	**G/A HET**
ABO001.22	**O**	***O01(ABO*O.01.01)/ O01(ABO*O.01.01)***	G	G
ABO001.23	**O**	***O01(ABO*O.01.01)/O26***	**C**	**G**
ABO001.24	**O**	***O01(ABO*O.01.01)/ O01(ABO*O.01.01)***	G	G
ABO001.25	**O**	***O01(ABO*O.01.01)/ O01(ABO*O.01.01)***	G	G
ABO001.26	**O**	***O02(ABO*O.01.02)/ O02(ABO*O.01.02)***	**C**	**G**
ABO001.27	**B**	***B101(ABO*B.01)/ O01(ABO*O.01.01)***	G	**G**
ABO001.28	**B**	***B101(ABO*B.01)/ O01(ABO*O.01.01)***	C/G HET	**G**
ABO001.29	**B**	***B101(ABO*B.01)/ O01(ABO*O.01.01)***	C/G HET	**G**
ABO001.30	**B**	***B101(ABO*B.01)/ O01(ABO*O.01.01)***	G	**G**
ABO001.31	**B**	***B101(ABO*B.01)/ O01(ABO*O.01.01)***	G	**G**
ABO001.32	**B**	***B101(ABO*B.01)/ O01(ABO*O.01.01)***	G	**G**
ABO001.33	**B**	***B101(ABO*B.01)/ O63***	C/G HET	**G**
ABO001.34	**B**	***B101(ABO*B.01)/ O01(ABO*O.01.01)***	G	**G**
ABO001.35	**B**	***B101(ABO*B.01)/ O75***	C/G HET	**G**
ABO001.36	**B**	***B101(ABO*B.01)/ O01(ABO*O.01.01)***	C/G HET	**G**
ABO001.37	**B**	***B101(ABO*B.01)/ O02(ABO*O.01.02)***	C/G HET	**G**
ABO001.38	**B**	***B101(ABO*B.01)/ O01(ABO*O.01.01)***	G	**G**
ABO001.39	**B**	***B101(ABO*B.01)/ O02(ABO*O.01.02)***	C/G HET	**G**
ABO001.40	**A_2B**	***A201(ABO*A2.01)/ B101(ABO*B.01)***	**G**	**G**
ABO001.41	**A_2B**	***A201(ABO*A2.01)/ B101(ABO*B.01)***	**G**	**G**
ABO001.42	**A_2B**	***A201(ABO*A2.01)/ B101(ABO*B.01)***	**G**	**G**
ABO001.43	**A_2B**	***A201(ABO*A2.01)/ B101(ABO*B.01)***	**G**	**G**
ABO001.44	**A_2B**	***A201(ABO*A2.01)/ B101(ABO*B.01)***	**G**	**G**
ABO001.45	**A_2B**	***A201(ABO*A2.01)/ B101(ABO*B.01)***	C/A HET	**G/A HET**
ABO001.46	**A_2B**	***A201(ABO*A2.01)/ B101(ABO*B.01)***	G	**G**
ABO001.47	**A_1B**	***A101(ABO*A1.01)/ B101(ABO*B.01)***	C/A HET	**G/A HET**

Table 5.9 The analysis of minisattelite repeats 3.8kb upstream of the start codon of *ABO*.

Four repeats were covered by the NGS data from all *ABO* samples with different ABO phenotypes. There was a SNP in nt. 41 in the first and fourth repeats, at chromosomal locations chr9: 136154433* and 136154304**, respectively.

The repeat sequence is ACCCCAGCCAATAGGGGAAGGACACAGAAACAGAAACTGC**G**TT, where the underlined **(G)** is nucleotide 41.

5.3.7 Validity of NGS data

The variant analysis from the NGS data is valid and need no further validation due to the high quality, coverage depth and all SNPs were reported to the dbSNP database. However, at the beginning this project (first NGS *ABO* run), the SNPs in exons 3, 4 and 6 were investigated and confirmed by Sanger sequencing. Primers were designed to amplify around SNPs in exon 3 (103G/A, 106G/A), exon 4 (188G/A, 189C/T) and exon 6 (297A/G). These SNPs were selected for confirming the NGS of *ABO* run; 103 G/A was a rare SNP and the rest were randomly selected. The primers used for amplification of these SNPs are listed in Table 2.13; with thermocycling conditions described in Table 2.14. The Sanger sequencing results matched the NGS data (Figure 5.14).

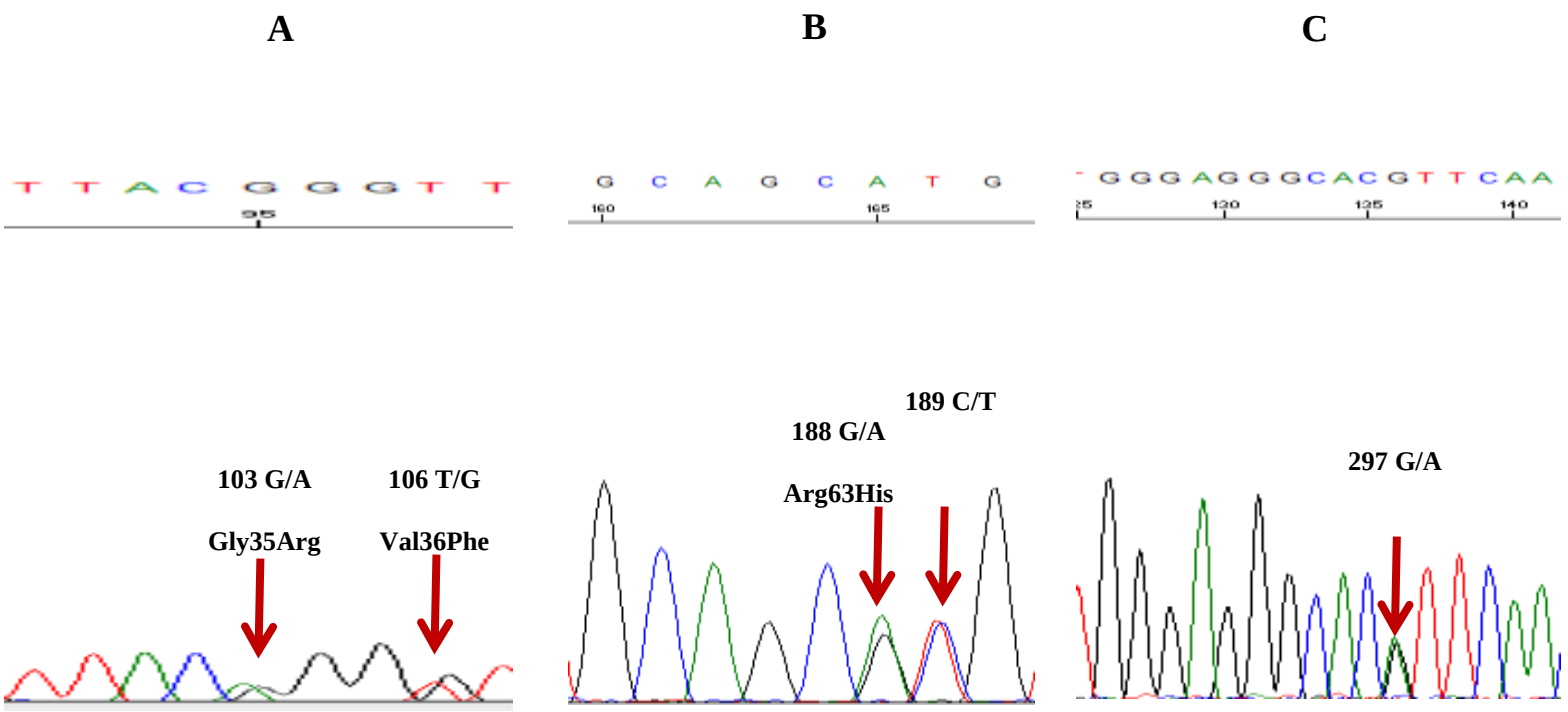


Figure 5.14 Sanger sequencing of a number of SNPs to confirm NGS data of one sample.

Several SNPs identified by NGS genotyping of samples of O phenotype (ABO001.12 shown here) were investigated by Sanger sequencing for confirmation of the integrity of NGS SNP data. All SNPs were heterozygous and matched NGS genotyping data. The reference sequence , particularly nucleotide locations 106, 188, 189 and 297 resembled the *O02* genotype.

5.4 Discussion

5.4.1 *ABO* NGS genotyping with LR-PCR

ABO genotyping is a valuable, complementary approach to serology, especially to elucidate discrepancies, and overall leading to better determination of the ABO status of donors and patients to ensure safer clinical transfusion. The causes of the discrepancies may be due to weak expression alleles or hybrid alleles. A_x alleles may be mistyped as O group due to low reactivity with anti-A, while the serum contains anti-A1 (Olsson et al., 2001, Daniels, 2013). Hosseini-Maaf et al. (2005) illustrated an example of solving the serological discrepancy by genotyping in 16 samples, which were initially heterozygous for O^2 and O^1 alleles. However, they manifested discrepancies which led to serological mistyping, one case of which was typed as A_{weak} due to weak reactivity with anti-A/B and an absence of anti-A. DNA sequencing analysis of this individual revealed the conventional *O02* genotype, with O^2-like polymorphisms and additional SNPs 649C>T and 689G>A – which were novel at the time of the study (Hosseini-Maaf et al., 2005). The majority of *ABO* genotyping studies have relied on methods such as sequence-specific primer (SSP) PCR, the output of which are insufficient for extended genotyping of transfusion units that are high in demand (Flegel et al., 2015). Moreover, other high-throughput approaches, such as microarrays, are only capable of revealing known *ABO* alleles (Avent et al., 2007, Avent et al., 2015). As a result, novel alleles are likely to be missed, which are increasing significantly as reported in the study of Lang et al. (2016), which used the NGS approach to genotype *ABO* samples and estimate the frequency of *ABO* alleles in a German population. Although only exons 6 and 7 were sequenced in Lang's study, a suggested 287 novel alleles were found and were not previously described in the BGMUT (which in turn reported an additional 95 new alleles four years since the publication of (Patnaik et al., 2012). A considerable number of *ABO* genotyping studies focused only on exons 6 and 7 (Lang et al., 2016, Taki and

Kibayashi, 2014, Hosoi, 2008), as these exons are thought to carry the majority of the *ABO* coding sequence (Daniels, 2013). However, it is argued that genotyping all exons and introns allows for more accurate assessment and prediction of the ABO phenotype and thus promotes safer transfusion practice and forensic application; this is especially relevant to subgroups or other *ABO* alleles, such as *O02,* new alleles or hybrid alleles carrying various polymorphisms across the *ABO* gene (Huh et al., 2011). In the exon 6 and 7 restricted Lang et al (2016) study, it has been stated that high resolution genotyping would have been obtained if the rest of the gene was sequenced since multiple alleles were not covered, such as *ABO**A_w*.13.01.1* which differed from the *ABO*A1.01* by a SNP in exon 1 (Lang et al., 2016).

Here, the high-throughput NGS approach was used to extensively genotype the entire *ABO* gene, including all 7 exons, introns and the flanking regions (up to 5kb upstream). This protocol ensures that all existing and potential novel polymorphisms are covered, enabling maximally thorough analysis of *ABO* molecular biology and, thus, better insight into the phenotype. The NGS protocol here was coupled with LR-PCR to yield 4 overlapping amplicons, which were used to amplify the entire *ABO* gene. Using a small number of primer pairs for PCR amplification is more time- and cost-effective and carries a lower risk of error and is less labour intensive. In a previous study by Huh et al. (2011), there were 31 sequencing reactions in total that required more than 17 sequencing primer pairs, in addition to two long amplicons, to genotype the *ABO* gene with the exclusion of intron 1 (~13 kb). Here, the entire *ABO* gene, including flanking regions, was successfully amplified by NGS of four LR-PCR-produced amplicons. Using only four amplicons increased the likelihood of successful amplification, as there would likely be fewer polymorphisms surrounding primer binding areas; this is especially true for highly polymorphic areas, such as in exon 7 of *ABO*. This was encountered in a study by Fichou et al. (2014), who used NGS Ampliseq^{TM} technology

(in which multiplex PCR is used) to sequence the coding regions of genes and intron-exon boundaries of 15 blood group systems. The amplification of several areas including exons 1-4 was unsuccessful, most likely as a result of homopolymers and GC-rich sequences, especially around exon 1 and upstream regions, which would have required nonstandard amplification conditions (Fichou et al., 2014).

5.4.2 Quality control of NGS data

The quality of the NGS data was assessed via FastQC, along with the coverage depth analysis, based on the Phred score. The quality of data was found to be high (>99% base call accuracy, with a 1 in 1000 probability of an incorrect base call) and there was a good mean coverage depth of 650X, which significantly exceeds that suggested to be sufficient for accurate variant calling (30X) (Tilley and Grimsley, 2014). This increases confidence in the NGS data for downstream analysis (Sims et al., 2014). The high coverage seen here is thought to be attributed to only using part of the 316^{TM} chip capacity for each *ABO* experiment, as 111 samples can be simultaneously sequenced for the entire *ABO* gene (~ 36kb amplicon size) by this chip at a coverage of 50X. The output can be further increased using higher capacity chips and would decrease the sequencing cost of the entire *ABO* gene. The library preparation cost for one *ABO* sample was ~£69, while the cost of sequencing run of one sample was ~£4. Using higher capacity chips, for example 318^{TM}, would allow sequencing of 555 *ABO* samples at once at a cost of ~£1. As a result, the high throughput capability of the NGS would eventually lead to a drop in the costs of blood group gene genotyping and might promote its use in more clinical laboratories in the future. The impact of implementing the NGS approach on reducing the costs of genotyping has been pointed out, as in a study of typing HLA alleles by NGS high-throughput platform (Illumina MiSeq) that led to a reduction of the costs by more than 50% compared to Sanger sequencing; moreover, the NGS data was in complete concordance with results obtained from

Sanger sequencing (Lange et al., 2014).

5.4.3 Validity of *ABO* NGS

The validity of NGS genotyping of blood groups here has already been discussed in the previous two chapters (*FY* 3.4.3 and *JK* 4.4.3). Similarly, the genotyping data quality for *ABO* was high here, as assessed by FastQC and coverage depth. In addition, there was a complete concordance of the phenotype from the serological approach with that suggested from the NGS, in fact the NGS provided more detailed and accurate phenotype (for example, the serological phenotype was AB while that of the NGS was A_2B). Moreover, the polymorphisms found were already reported in the dbRBC database. Nevertheless, despite these parameters that illustrated the validity of the *ABO* NGS data, Sanger sequencing was used for a number of SNPs, which were found at the beginning of this study, for their confirmation and this data was in agreement with NGS output (section 5.3.7).

5.4.4 NGS *ABO* library

The first step to constructing *ABO* libraries for NGS genotyping of the whole gene was PCR amplification of four long amplicons. Some difficulties were encountered when amplifying the amplicon covering the region upstream-exon 1 and part of intron 1 in samples of various phenotype and thus genotype (section 5.3.4.2), which required another primer pair to be used. Visualisation analysis of amplicons that failed PCR amplification but were sequenced regardless revealed SNPs surrounding primer binding sequences, which may have perturbed amplification as these SNPs were not found in samples that were successfully amplified (Table 5.5). Correlation of these SNPs with *ABO* alleles was not confirmed, although SNP 136149500 A>G was found to be homozygous in the *A201/A201* sample. Analysing more samples, especially ones with homozygous alleles, would allow correlation of SNPs with the alleles; moreover, further

analysis, such as Sanger sequencing, of the primer binding sites would be useful, for instance, to explain the lack of coverage at the first nucleotide attaching the reverse 1A primer in our study. It is advised to avoid these SNPs when designing primers, which was the strategy for the design of primer pair 1B here. Another possible reason for the failure of amplification may be a GC-rich area around exon 1, which may require nonstandard amplification conditions as has been mentioned (Fichou et al., 2014). It is worth paying attention to this area in future *ABO* studies when optimising amplification conditions. With regard to amplicon 2, which mostly covers intron 1, there were a few samples that failed to be amplified or very small amplification was noticed (faint bands); thus, the corresponding area was not covered by NGS. It is unclear why this region displayed amplification problems, as visualisation analysis showed no signs of polymorphisms around the primer binding areas (which were covered by NGS), yet the middle of the amplicon was not sequenced (covered). Possible reasons could be a high GC-content which needs optimisation of PCR conditions; however, since the majority of samples were amplified, another reason could be the presence of homopolymer sequences across the *ABO* gene (repeats of the same base in the sequence) as previously suggested by Fichou et al. (2014). These homopolymers could affect amplification and also NGS sequencing by the Ion PGM™ platform, leading to errors (Loman et al., 2012, Fichou et al., 2014, Bragg et al., 2013, Quail et al., 2012).

With regard to the *ABO* library processing, the purifcation and size selection was conducted by SPRIselect® reagent kit, which is cost effective and less time consuming compared to Pippin Prep™ instrument (section 2.2.4.10).

5.4.5 NGS data analysis (*ABO* reference sequence)

According to the ISBT and NCBI databases, the *A101* allele is a consensus reference sequence against which the variants are annotated (Reid et al., 2012). However, this is likely to be developed by the NCBI database project (RefSeqGene) used for

conventional annotations which is stated to be developed with the consultation of the gene-specific experts (NCBI website). It was found here that the hg 19 reference sequence of the *ABO* resembled a combination of two alleles (*O01* and *O02*) (see section 5.3.4.1), with a deletion (261delG) in the reference sequence both in hg19 and hg38. This issue with the hg19 *ABO* reference sequence was previously noted by Lane et al. (2016), whose study involved red cell antigen prediction from whole genome sequencing (Lane et al., 2016). As a result, a G insertion would be in samples containing *A* and *B* alleles (and this has been observed), thus causing downstream nucleotide changes and resulting amino acid location shift that needs to be noted. However, here NGS variant analysis against the NCBI RefSeqGene (*A101*) was not feasible as the chromosomal locations of the polymorphisms may not be accurate as the sequence has different coordinates to that of hg19 or hg38. Consequently, NGS data was aligned against the hg 19 reference sequence for several reasons: to be consistent throughout this project (*FY* and *JK* was analysed against hg19); it has been utilised by previous NGS studies (Rieneck et al., 2013, Fichou et al., 2014, Lane et al., 2016); and to be consistent with data analysis software (Ion Reporter™, most updated version), which is based on hg19. Moreover, it is worth noting that the visualisation analysis of *ABO* was challenging due to its anti-sense direction (section 5.3.4.1), which may lead to false annotation. Nevertheless, Ion Reporter™ reports SNPs in the sense direction and this issue could also be manually resolved by flipping the sequence.

5.4.6 Genotyping of the *ABO* gene

5.4.6.1 Polymorphisms in exons

NGS genotyping of the 47 *ABO* samples, all of which were serologically phenotyped, revealed numerous exonic and intronic polymorphisms which accounted for different *ABO* alleles (Table 5.6 and 5.7). NGS also provided more detail on sample phenotype (e.g. A_1 and A_2) than serology. Genotyping of samples of A phenotype revealed a comparable frequency of A_2 phenotype compared with those previously reported. Here, the A_2 phenotype frequency among 11 samples with A phenotype was 27% with 5 *A2* haplotype. This is a slightly higher frequency than that described in a study that illustrated the distribution *ABO* allele frequency in England, with the A_2 phenotype representing 22% of the A phenotype samples (Ikin et al., 1939, Daniels, 2013). The possible reason for this increase could be due to the small number of samples of A phenotype being analysed here, as the study by Ikin et al. (1939) analysed 1546 samples of A phenotype, upon which the frequency was provided. On the other hand, the frequency of the A_2 phenotype among the samples analysed in this project was similar to that found in a previous study that genotyped 34 A phenotype samples from Birmingham, England (finding that ~70% and 30% were A_1 and A_2, respectively) (Procter et al., 1997). Likewise, in the same study, the ABO blood group genotyping of A phenotype samples (63/146) from random renal donors revealed a comparable frequency (76% and 24% for A_1 and A_2, respectively).

With regard to samples of O phenotype, NGS identified rare polymorphisms that generated *O* alleles (section 5.3.6.1 and Table 5.6) previously reported in publications or in the BGMUT database. One of the *O* alleles identified here is suggested to be novel (0% frequency in Europe) and carries a missense SNP (482 G>A, Arg161His) in exon 7. Although this SNP has been reported and validated by the NCBI dbSNP database, it has

yet to be allocated to an *ABO* allele listed in BGMUT. For samples of B phenotype, there was complete concordance between NGS data and serology. Regarding the frequency of the samples with the AB phenotype, samples with A_2B phenotype showed higher frequency than A_1B phenotype. According to genotyping data, the majority of the samples (7/8; 87.5%) were of A_2B phenotype – a significantly higher frequency than reported in English donors (~19.5%, 22 out of 113 donors with AB phenotype) by Iken et al. (1939) (Ikin et al., 1939, Daniels, 2013). The possible reason for this finding may be the small number of samples (8 AB samples) used here, which may have resulted in invalid representation of allele frequency in a population compared to the study of Iken et al (1939), in which the analysis of 113 A phenotype individuals was conducted. The results here are therefore likely a coincidence, the 7 randomly selected samples happened to carry the A_2B antigens. The A_2B phenotype samples were observed at a higher frequency than A_1B phenotype samples in an early study by Procter et al. (1997), in which 63% of the samples of AB phenotype (8 AB phenotype samples/146 total samples) from renal donors were A_2B and 37% were A_1B. Accordingly, sequencing large number of samples would provide better illustration of the phenotypes and thus allele frequencies in different populations, which is feasible by the high-throughput NGS capability. Lastly, the majority of our samples were heterozygous, in agreement with previous reports (Daniels, 2013).

5.4.6.2 Polymorphisms in introns and upstream regions

The *ABO* NGS genotyping approach used here enabled analysis of all existing polymorphisms across the gene, including those in introns and upstream areas. There are several suggested benefits to analysing intronic polymorphisms, namely it yields information on gene and allele evolution and hybrid alleles. The latter was observed here with respect to the hg 19 *ABO* reference allele (Table 5.8) and in a previous study of the *Bw26* allele, which revealed a hybrid sequence of *O03* and *B101* alleles in both introns and exons (Thuresson et al., 2012), thereby providing more accurate genotyping data that might help in the forensic analysis of the *ABO* genotyping (Huh et al., 2011, Avent et al., 2015). Here, all introns were genotyped, including intron 1 (~13 kb) which was not covered in the previous entire *ABO* genotyping study (Huh et al., 2011). The analysis of intron 1 is suggested to be beneficial for studying the expression of ABO antigens on the surface of red cells, as it has been reported to contain sequences (5.6-6.1 kb from the translation initiation site) suggested to be erythroid specific enhancer elements that plays a role in GATA-1 transcription factor binding. The deletion of which was previously shown to be associated with a B_m phenotype with no erythroid B antigen expression on red cells while B antigen found in saliva (Sano et al., 2012). However, it is worth noting to consider intronic SNPs when designing primers as they may affect the success of amplification, as seen with *ABO* amplicons 1A and 1B here (section 5.3.4.2), and would result in allelic dropout.

A high number of intronic SNPs was found here across different *ABO* alleles; however, not all of these SNPs underwent correlation analysis with alleles, due to time constraints since it was done manually (section 5.3.6.2). The attempt to establish a specific set of intronic polymorphisms, as done with the *JK* alleles (section 4.4.4.2), was challenging due to the fact that the majority of samples were heterozygous with an *O* allele, despite the phenotype similarity. Thus, future NGS sequencing studies should involve more

ABO samples with homozygous alleles, such as *A101/A101*, to study correlation of alleles with intronic polymorphisms. This would be important, as polymorphisms in *ABO* introns have previously shown use as allelic markers along with exonic SNPs (Hosseini-Maaf et al., 2003). Otherwise, the technology of the single molecule sequencing may circumvent this issue with regard to the assignment of polymorphisms to corresponding alleles but it was beyond the remit of this project (see section 6.4). This was observed here during analysis of samples homozygous for the *O* allele; although, the number of these samples was small. Nevertheless, there were various instances of observed complete correlations between intronic SNPs and alleles, such as the 13bp insertion in intron 4 of the *O02* allele and from exons 3 to 7, including the introns between, with *O01* and *O02* alleles were noticed (Table 5.8). A complete correlation of alleles with SNPs in intron 1 was not seen here, however, which could be due to the limited number of intronic SNPs analysed in this area. As the correlation of intronic SNPs analysis was carried out manually, the comprehensive intronic SNPs analysis (more than 70 SNPs) in the 13kb intron 1, was challenging and time consuming; therefore, more efficient analysis in the future may lead to better correlation data. Interestingly, analysis of rare *O* alleles, which resembled either *O01* (*O26* and *O28*) or *O02* (*O63, O68, O73* and *O75*), was suggestive of allele evolution.

As in the analysis of the NGS data regarding the polymorphisms in the exons, NGS analysis of introns illustrates zygosity, which is used to evaluate the specificity and correlation to alleles. This was seen in the analysis of the homozygous *O* alleles analysis with the *O01*/*O75*, which was used as heterozygous sample. However, in the samples with AB phenotype there was an unclear situation in that none sense zygosity pattern was seen around the area covered by amplicon 2, particularly around the 22 bp deletion chr9:136145994-chr9:136146015 (Table 5.7) in intron 1. Instead of heterozygous, AB phenotype samples were homozygous difference to the reference in that area, whereas

no observed change was seen in all B and A_1, although those with A_2 phenotype were similar to that in the AB phenotype samples. There were no visible polymorphisms that may have led to allelic dropout (which is more likely in the *A101* and *B* alleles) seen here when analysing primer binding sites of amplicon 2. Thus, the reasons for unsuccessful amplification of this amplicon here are unclear and it may be worth revisiting the design of primer pairs and amplification conditions in addition to sequence and analyse more samples with AB phenotype in the future.

Moreover, it is worth conducting a more thorough investigation of intron 1 in the future. With regard to the analysis of the upstream area, there was no correlation between the number of repeats of the CBF/NF-Y transcription factor binding site and the *ABO* alleles, as all four copies were present in all alleles. This finding contradicts the suggested description of the alleles correlation with the number of copies carried in which all (4 copies) are found in *A2, B,* O^1 alleles while one copy presented in the A^1 and O^2 alleles. This suggestion was depending on the PCR-product fragment size and number of bands that was amplified followed by sequencing (Irshaid et al., 1999). Therefore, these findings merit further investigation in the future, as there are conflicting views in the literature on the effect of variations in this area on the ABO antigen expression, especially in homozygous *ABO* samples. In 2007 a study suggested a reduced B antigen expression due to variations in the CBF/NF-Y region (Seltsam et al., 2007). Another study contradicts this assumption by arguing that the number of repeats does not affect transcription, which was normal, in *Bw26* alleles, whereas reduced expression occurs as a result of the SNP 53G>T (Arg18Leu) (Thuresson et al., 2012).

In conclusion, used together with LR-PCR, the NGS approach enabled comprehensive genotyping of the *ABO* gene. In this study, the entire *ABO* gene plus outer regions were sequenced, which enables identification of all known, rare and novel polymorphisms across the gene and therefore permits the study of new alleles. The high-throughput and discovery capabilities of NGS genotyping opens an avenue for a variety of future investigations with regard to the *ABO* gene; these may help elucidate unusual cases, such as hybrid and low frequency alleles, and solve any discrepancies between the *ABO* genotype and serological phenotype. It is, therefore, concluded here that NGS is a potentially reliable and feasible platform for genotyping complex blood groups such as *ABO*. In addition, as NGS technology and data analysis software continually improve, it is likely that this tool may in the near future become commonplace in the clinic for the purpose of genotyping patients and donors for safe transfusion practice.

Chapter 6

General Discussion and Conclusion

The determination of some blood groups in individuals is essential for the safety of blood transfusion. Mismatched exposure to blood group antigens leads to alloimmunisation that may result in transfusion reactions. Correctly matched blood units are crucial for patients, especially those who require multiple blood transfusions for survival, such as those with sickle cell disease (SCD). The occurrence of alloimmunisation is frequent among SCD patients receiving blood units that do not undergo extended blood group antigens matching (Chou and Westhoff, 2011). As a consequence, the possibility of HTR, as well as the difficulty in finding compatible RBC units for subsequent transfusion, which might risk the recipient's life, pose great challenges in the successful management of such patients (Bakanay et al., 2013). Also, Alloimmunisation occurs in pregnancy in cases, such as fetomaternal haemerhage. Comprehensive and extended blood group antigen typing is therefore crucial and has, in fact, been shown to reduce the incidence of alloimmunisation (Chou and Westhoff, 2011). Serological typing might be insufficient for accurately defining blood group phenotypes in multiply transfused patients, especially if they have recently received transfusions, owing to the possibility of persistence of circulatory RBCs of the donor (Castilho et al., 2002). In addition, serological typing for an extended range of clinically significant blood group antigens would be expensive, particularly if applied to cover large numbers of donors and recipients (an example is provided in section 6.6). BGG is capable of overcoming the limitations of serology. Nevertheless, the commercially available BGG platforms have certain drawbacks, such as low throughput (except micro-array-based technology) and inability to define novel or rare alleles which are not included in the assay design. Accordingly, the emerging alleles, including those

affecting the phenotype that could be clinically significant, cannot be detected with these platforms. The NGS approach circumvents these issues by not only addressing the large demand on BGG applications with the ability for high-throughput genotyping but also making available the discovery mode based on its sequencing-based genotyping that defines all polymorphisms irrespective of their mechanisms, describing all alleles (known, rare and novel) found in samples (Avent et al., 2015). In this PhD project, an NGS approach was coupled with LR-PCR to assess the feasibility of NGS in genotyping blood groups.

6.1 NGS with LR-PCR

The ability of NGS-based genotyping of blood groups to define rare and novel alleles allows more comprehensive and accurate prediction and representation of the phenotype than currently possible. This makes this approach superior to those platforms based on predefined SNPs, which might falsely predict a positive phenotype encoded by rare or novel silencing or weakening alleles that are not included in the assay. One example of this situation is illustrated in (section 3.4.5), in which a discordant of the predicted phenotype by the HEA BeadChip assay Fy(b+) with that of serology Fy(b-). A sequencing approach confirmed a missense mutation in exon 2 895G>A (Ala299Thr), which was suggted to be a new silencing SNP that was not covered in the array based platform (Westoff et al., 2014).

The NGS approach was coupled with the LR-PCR method, rather than the AmpliSeq^{TM} technology, to develop a protocol for comprehensive BGG. The LR-PCR approach was used over AmpliSeq^{TM} technology for several reasons. The AmpliSeq^{TM} technology used with the Ion Torrent NGS platform is based on a customised panel of multiplex primers provided by the company by using Ion Ampliseq^{TM} Designer to generate a large number of amplicons of targeted areas in various blood group genes. Therefore, despite

its ability for high-throughput and sequencing-based genotyping that allows discovery of novel polymorphisms, this approach raises some constraints due to the fact that only targeted regions are genotyped in addition to the high number of amplicons needed. The Ion PGM™ with AmpliSeq™ technology has been previously used to genotype (only the exons) of 11 blood group genes (Halawani, 2015). It has been pointed out that the NGS with AmpliSeq™ technology protocol was simpler in terms of the library preparation, in which no fragmentation nor purification was needed. However, a large number of primer pairs were designed to cover the targeted regions in that study; for example, 6, 9 and 11 amplicons were needed to cover only the exons of *FY*, *JK* and *ABO*, respectively. As a result, amplicon coverage might be affected by the failure of amplification targets of a number of primers due to polymorphisms at the primer-binding sites. In addition, the designed primers might not have sufficient and accurate specificity. This, in fact, was evident since several regions of blood groups were not covered; for example, two regions of exon 7 of *ABO* were missed. In addition, the designed primers provided in the panel failed to distinguish the homology between the *RHD* and *RHCE* genes by misaligning a SNP for *RHCE* to *RHD*, which was suggested to be due to the lack of specificity in the primer design. The latter issue was addressed by designing gene specific primers with the LR-PCR approach, coupled with NGS (Halawani, 2015, Avent et al., 2015). Similar issues were encountered in the 2014 study by Fichou et al, in which AmpliSeq™ technology was used with a designed panel to generate 417 amplicons (16, 31 and 16 for *FY*, *JK* and *ABO*, respectively) for genotyping the coding and untranslated regions of 18 blood group genes involved in 15 blood group systems. The authors mentioned difficulties in genotyping with only 86% of the coding regions being covered, especially for the *ABO* gene, where several exons were not covered (1 and 4), whilst irrelevant sequences were obtained for the remaining exons. In addition, the designed panel of primers failed to cover and distinguish

homologous genes, for example *RHD* and *RHCE*. This has also been pointed out to have been solved by LR-PCR with specific primers (Fichou et al., 2014). Accordingly, NGS with LR-PCR was used here to sequence the complete gene, including splice sites, introns and flanking regions, which provides an extensive genotyping and analysis of the blood group genes.

To our knowledge, this is the first NGS-based genotyping with LR-PCR of the blood group genes *FY*, *JK* and *ABO*, although LR-PCR was used before in a *FY* sequencing study but not with NGS technology (Schmid et al., 2012). Since complete coverage of the sequence across the gene can be visualised by the IGV system, an overview of polymorphisms in critical areas, such as exons and splice sites, can be rapidly obtained. Further investigation will then be considered, depending on the knowledge of the critical molecular locations of the genes. Polymorphisms across the genes were analysed, starting with those in the exons.

6.1.1 NGS data quality

The overall quality of the NGS data from genotyping all three genes was high. According to the Phred score, the base call accuracy was more than 99%, with a 1 in 1000 probability of an incorrect base call. In addition, the coverage depth (for example, 5600x for *FY* genotyping), which is suggested to increase the confidence for variant analysis results, was significantly higher than pointed out to be sufficient (30x) (Tilley and Grimsley, 2014).

6.2 Polymorphisms in Exons

The entire *FY* gene was extensively genotyped by NGS with LR-PCR, using only one single amplicon to cover exons, intron and promoter region. All exonic polymorphisms in the samples were revealed. The NGS genotyping data matched and confirmed the phenotypes of *FY* alleles *FY*A*, *FY*B* and *FY*02*(*Null*) in terms of the crucial SNPs

previously described (Reid et al., 2012). For example, the crucial SNP for *FY*A/FY*B* (125A>G) in exon 2 encoding for Gly42Asp was consistent with the provided Fy^a/Fy^b phenotype. The 298G>A encoding Ala100Thr aa change, which occurs on the *FY*B* allele background since none of *FY*A/FY*A* samples showed this SNP, was commonly found in our samples (16/53); as stated before, this alone is not expected to affect the Fy^b expression (Olsson et al., 1998).

The NGS genotyping of the *JK* was feasible and allowed extensive analysis across the gene. NGS genotyping data of the key *JK* allele SNP (*JK*A/JK*B* 838G>A) was in concordance with the provided serological phenotype. The SNP 130G>A, which was described to be specific to the *JK*01W.01* allele associated with a weaken expression of Jk^a, was found to be heterozygous in 7 *JK*A/ JK*01W.01*, 2 in *JK*B/ JK*01W.01* and 1 homozygous *JK*01W.01/ JK*01W.01* samples. With regard to the expression of Jk^a, the phenotype in the samples appeared not be affected by the SNP. Although it could be argued that even weak expression is recorded as positive with the system used by the sample provider, it was found that two samples with this SNP (homozygous and heterozygous, respectively) reacted strongly in the haemoagglutination assay (Wester et al., 2011). Therefore, it was speculated that this SNP 130G>A might not be the only factor in reducing Jk^a expression; other polymorphisms located elsewhere across the gene that were not observed in Wester's et al (2011) study might also play a role in bringing about the reduced expression. The approach of NGS and LR-PCR used in the present project might be a great candidate for the investigation in determining these polymorphisms since the entire gene and flanking region is covered by only three amplicons. Accordingly, samples with discrepant phenotype or weak expression (weak Jk^a expression in this case) can be sequenced, and critical areas such as splice sites, introns and upstream region, which contain erythroid-specific GATA-1 transcription factor binding sites (section 1.7.1), can be analysed. The ability of high-throughput

screening allows for the analysis of a large number of such samples to define the weakening allele-specific polymorphisms. The visualisation of the NGS sequence across the gene allows rapid identification of polymorphisms within significant locations. The SNP 810G>A locates in the second last nucleotide in exon 8 (exon 8/intron 8 boundary), which might raise the concern about affecting the splicing. In fact, this has been pointed out in a study that suggested the silencing effect of this SNP on the Jk^b expression resulting from the association with a novel *JK*B Null* allele (Henny et al., 2014). However, all samples carrying this SNP showed no effect on the expression of the Jk^b (phenotype Jk b+), in addition to the confirmation obtained from the cDNA Sanger sequencing that showed a correct splice out of intron 8.

The LR-PCR coupled with NGS approach was also feasible to comprehensively genotype *ABO*, including all 7 exons, introns and flanking regions. This is different from other studies that either sequenced some of the exons or neglected intron 1, possibly due to its large size (see chapter 5). This allows accurate elucidation of the molecular basis of the various *ABO* alleles, identifying rare, novel polymorphisms or hybrid alleles that would enable better prediction of the phenotype. An example of the benefits of genotyping the entire *ABO* gene and not only exons 6 and 7 is the definition of the *O02* allele since it carries polymorphisms in different exons, including 3 and 4, across the gene. Several rare alleles and one novel *O* allele were identified using the NGS and LR-PCR approach with only four long amplicons. Using a small number of primer pairs for PCR amplification is time- and cost-effective and possibly carries a low risk of error and is less labour intensive, especially if accompanied with robotic automated PCR set up. In addition, the likelihood for successful PCR amplification with the use of a small number of primer pairs is greater compared to the case with the use of a large number of primer pairs, which increases the chances of encountering polymorphisms at the binding sites (Mullins et al., 2007). This is particularly relevant in

the case of genes with a complex and highly polymorphic molecular basis such as *ABO*, which should be taken into consideration while designing primers. In addition, the upstream area/exon 1/intron 1 were found to be more complex for amplification than the rest; this necessitated the redesigning of another amplicon to address this issue, which has been reported to be problematic and might need nonstandard conditions for amplification (Fichou et al., 2014) (see chapter 5). Another benefit of the approach of sequencing the entire *ABO* gene by the NGS technology followed by comprehensive analysis and visualisation is the identification of hybrid alleles. Interestingly, the reference sequence from hg19 and hg38 were not only found to carry a 261delG polymorphism but also the sequence from hg19 resembled a combination of two alleles (*O01* and *O02*). This issue was also reported recently and should be taken into consideration during analysis in the future (Lane et al., 2016). The frequency of the predicted A_2B phenotype was found to be relatively high which could be inaccurate and requires larger samples to assign the frequency (section 5.4.6.1).The A_2B phenotype frequency among AB phenotype samples in this project was higher (87.5%) than A_1B samples, which was also seen by study that illustrated a frequency of 63% and 37% for A_2B and A_1B, respectively among (8 AB samples) (Procter et al., 1997). On the other hand, the frequency of A_2B was (~19.5%, 22 out of 113 donors with AB phenotype) in the study of (Ikin et al., 1939).

6.3 The Intronic Polymorphisms

The NGS approach applied here involved sequencing the introns and flanking regions to enable an extensive analysis of polymorphisms across the gene. The NGS data revealed a large number of intronic SNPs, for example up to 80 SNPs in the *JK*A/JK*B* samples. Although it might be argued that intronic polymorphisms, especially SNPs, might not affect the protein sequence and thus affect the expression of anitgens, several benefits have been obtained with the approach of analysing the intronic polymorphisms. It has

been suggested that the knowledge of molecular background, including information from introns and flanking regions, for weakening or silencing alleles is important to determine the causative factors of these alleles, where intronic/flanking region polymorphisms may play a major and/or cumulative role in terms of expression (Avent et al., 2015). For example, a deletion in intron 1 at (5.6–6.1 kb from the translation initiation site) is reported to abolish the expression of B antigen from the RBCs, which is suggested to be associated with B_m phenotype, while an intronic SNP 190C>T has been suggested to play a cumulative role in the reduced expression of Fy^b (Gassner et al., 2000, Sano et al., 2012)

In addition, intronic SNPs, along with those key allele defining SNPs in exons, might be used to distinguish between alleles more accurately. In fact, allele-specific intronic polymorphisms were described here after analysing homozygous samples, especially, those for *JK* alleles (*JK*A, JK*B* and *JK*01W* [*JK*01W.01*]), which might be considered the reference sequences for those alleles (see section 4.3.4.2, Table 4.5 and Figure 4.11). Consequently, variation from these sequences might indicate different alleles. Another example of the correlation of the intronic SNPs with the alleles was noted for the homozygous *O* alleles (Table 5.8).

Additionally, the intronic polymorphisms analysis approach may provide an insight into the allele evolution and hybrid alleles. An interesting example was the illustration of the *JK*01W.01* that appears to evolve from a hybrid *JK*A/JK*B* allele sequence, with the *JK*01W.01*-specifc polymorphisms being evolved on a hybrid *JK*A/JK*B* backbone (Table 4.5 and Figure 4.11). The hybrid sequence of the hg19 *ABO* reference sequence was also observed during the comprehensive analysis (exons, introns) of the *O* alleles (Table 5.8). In addition, various *O* alleles described here might be evolved from either *O01* (*O26* and *O28*) or *O02* (*O 63, O68, O73* and *O75*) alleles (section 5.3.6.2).

Moreover, cataloguing and acknowledging the intronic SNPs is recommended during the designing of primers, particularly those adjacent to key allele-specific SNPs. Such SNPs that might be around the primer-binding sites may hinder the amplification, which might cause an allelic dropout. One example of such a SNP is a C>T change found to be closely located (160 bp) upstream of the crucial *JK*A/JK*B* SNP (838G>A) (Section 4.3.4.2). In addition, intronic analysis revealed that the hg19 reference sequence might carry uncommon SNPs, with samples exhibiting a homozygous difference (section 3.4.4.2 and Figure 4.11).

Also, as is the case with exons, intronic SNPs might guide in the determination of zygosity, which might illustrate the correlation of polymorphisms to alleles, as seen in the case of the *O01/O75* sample that was used as the heterozygous sample (Table 5.8).

6.4 *Cis* or *Trans?* (Assignment of haplotype)

The current NGS platforms are designed to sequence small fragments of the DNA; for example, the Ion PGM™ can sequence 200–400 bp fragments. Accordingly, the allocation of a polymorphism, especially novel weakening or silencing, to a particular allele could be challenging. In other words, the definite assigment whether the polymorphism is in *cis* or *trans* to other polymorphisms may be difficult, since the fragmentation of the target (here, DNA) into small fragments may shatter those SNPs from the actual carrying allele. The LR-PCR approach is amenable for cloning into vectors for *cis/trans* assessment of novel mutations since vectors can hold inserts up to 10–42 kb (Casali. and Preston, 2003, Promega, 2016). However, allele-specifc LR-PCR might be applied to define alleles. Futhermore, advances in the sequencing technology have led to the emergence of the single molecule sequencing approach that allows single molecule sequencing of considerably long reads of 50 kb to 200 kb (or

theoretically any size) for PacBio and ONT MinION, respectively (Goodwin et al., 2016). Accordingly, this read length is possibly greater than all blood group genes and hence would be useful to assign the *cis/trans* location of polymorphisms, while eliminating the need for laborious gene cloning (Avent et al., 2015).

6.5 Data analysis

The software packages that were utilised for the NGS data analysis for genotyping and subsequent phenotype prediction are user friendly and do not require extensive knowledge of bioinformatics. On the other hand, fundamental and basic knowledge of the blood group systems, especially the molecular basis of the alleles and the correspondent blood group antigens, is crucial. Moreover, the depth and complexity of the analysis depends on the application needed, for instance whether only polymorphisms in coding regions or all those found across the entire gene, including introns and regulatory regions are needed.

In this investigation, the comprehensive data analysis was complex and time consuming since intronic polymorphisms were analysed, for example, to assign the correlation with blood group alleles and thus the effect on the antigenicity. This might not be feasible if NGS-based genotyping, which generates enormous number of data, would be applied routinely on genotyping donors and patients. To my knowledge, there is no dedicated software for the interpretation of NGS data for comprehensive BGG, although software for blood group phenotype prediction from NGS data has recently emerged but they might not provide extensive analysis (Giollo et al., 2015). Therefore, it would be worthwhile to establish a collaboration between experts in the blood transfusion science field and software developers to establish software packages that address the necessary applications. For example, the *JK* allele-specific coloured patterns (Table 4.5) described

here could be adopted to establish a rapid allele identification and phenotype prediction for scoring NGS data from samples automatically, in which polymorphisms, in both coding and non-coding areas, are rapidly called and interpreted. In addition, a rapid approach of interpreting the polymorphisms across the gene of different samples can be automatically and rapidly aligned and categorised among the samples and the reference sequence. The sheer volume of data generated from the NGS genotyping may raise concerns about the storage and analysing ability. This has prompted the development of cloud storage and computing that is suggested to address such concerns (Onsongo et al., 2014, Shanker, 2012, Bhuvaneshwar et al., 2015).

6.6 NGS costs

The NGS, with its large-scale and high-resolution genotyping, might replace the commercially available microarray-based platforms since it is not only cheaper but also precludes the constraints on the limited number of alleles that can be defined (Avent et al., 2015). With regard to serological testing, it has been reported that the costs of high-resolution testing, on average, were $195 (£149.49) per patient, while those with complex cases, such as autoimmune haemolytic anaemia, were on average of $1490 (£1142.24) per patient (Mazonson et al., 2014). On the other hand, the process of NGS extensive genotyping, from the library to the results of sequencing required £50.6, £67 and £73 per sample for *FY*, *JK* and *ABO*. The sequencing costs are dependent on the throughput and the reagents for each platform. The cost of the sequencing per sample drops significantly in higher throughput approaches (section 5.4.2). In addition to the reduction in the costs of the actual extensive genotyping, including all polymorphisms, high–throughput genotyping would eventually bring about a reduction in the cases of transfusion complications (since the frequency of alloimmunisation would decline) and preclude the unnecessary administration of prophylaxis to mothers, whose foetuses are not in danger of HDFN, although LR-PCR may not be feasible for cffDNA genotyping,

due to the size of the cffDNA (Avent, 2009, Finning et al., 2008).

6.7 Future advancements in NGS technology

NGS technology is undergoing continuous development in various aspects, for instance, speed, throughput, and automation of sample processing (such as DNA library preparation). These might help and simplify the implementation of large-scale routine BGG that is in high demand (Avent, 2009). An example of the approaches towards automation for the library preparation is SPRIworks system I by Beckman Coulter, which is designed for automated library preparation, in terms of adapter ligation, size selection and purification, which is attached with an automated nucleic acid extractor (SPRI-TE) (https://www.beckmancoulter.com). In addition, the Ion Torrent Chef™ system (from Thermo Fisher Scientific) was developed recently for automated template preparation and loading the samples into the chip automatically. It can also be involved in the library preparation of the AmpliSeq™ technology (https://www.thermofisher.com). These advancements, which are continously improving, might help reduce the labour intensive and inconsistency of processing a large number of samples for the high-throughput NGS platforms.

There have been several advancements in terms of the throughput that allows a significant number of samples and large targets, such as all blood group genes or whole genome to be simultaneously sequenced, which also reduces time and costs. Ultra-high-throughput platforms have been described. Examples of such platforms are Illumina HiSeq X and the single-molecule Oxford Nanopore PromethION with throughput capacity of up to 900 Gb and 4 Tb, respectively, while the cost has been reported to be only $1000 per genome with the former platform (Tilley and Grimsley, 2014, Goodwin et al., 2016). Accordingly, great benefits can be achieved, such as the possibility of sequencing donors and patients for all blood group genes simultaneously. One approach

could be the designing of an expanded panel with multiplex primers (providing improved specificity) to sequence all blood group genes, including exons, introns and regulatory regions, to enable accurate phenotype prediction. It has been pointed out that all blood group genes with regulatory regions comprise only a fraction of the human genome (~150 kb) (Tilley and Grimsley, 2014), whereby a significant number of samples can be simultaneously sequenced by continually developing throughput platforms. Furthermore, due to the continuous advancement in the throughput, the whole genome sequencing of individuals (donors and patients) might become the conventional approach applied from birth in the future. Accordingly, all the molecular information might be available for accurate phenotype prediction without the need for further expenses in investigation (Tilley and Grimsley, 2014). In addition, the high-throughput platforms, especially those for single molecule sequencing, might provide more accurate and refined molecular bases of the blood groups alleles and allow better prediction of phenotypes.

6.8 Future work

With regard to *ABO* analysis, more homozygous alleles (for example *A101/A101*) may be sequenced in the future for correlation assignment of polymorphisms with alleles. In addition, more samples with AB phenotypes could be sequenced to assign the frequency of the A_2B phenotypes since the number of samples in this investigation may have been inadequate for frequency assessment. Furthermore, the area covered by amplicon 2 (in intron 1) of the *ABO* gene might be revisited in terms of redesigning the primers and sequencing more AB samples to investigate the causative factors for the unusual observation (section 5.4.6.2). Moreover, the upstream region that includes the CBF/NF-Y transcription factor binding site might be revisited by other methods, for example, Sanger sequencing to investigate the number of copies of the (43-bp repeats) associated with *ABO* alleles. Furthermore, with the advances in the NGS data analysis software packages and the development of an extensive BGG specific software, a thorough analysis and scoring of the intronic polymorphisms, especially that in intron 1, can be accomplished in an easier and possibly automated fashion, considering an increase in the number of samples analysed in the future. With regard to *JK*, *FY* and *ABO* genotyping, samples with discrepant or weak phenotypes can be genotyped by LR-PCR to thoroughly investigate the polymorphisms across the genes, including flanking regions, and to investigate the causative factors, especially if encoded by rare or novel alleles with unknown molecular bases. This is because the NGS-based genotyping has no constraints in terms of only known SNPs. Moreover, single molecule platforms might be utislied in order to confirm the *cis/trans* location of a polymorphism to an allele.

6.8.1 Allele classification system

The comprehensive genotyping analysis (including the analysis of intronic polymorphisms) of homozygous samples defined the *JK* allele specific-patterns, thus suggesting reference sequences for *JK*A*, *JK*B* and *JK*01W.01*. Samples with sequence patterns that are different from the suggested reference sequence are considered different alleles. Accordingly, it is essential that a classification system for allele terminology is devised. Figure 6.1 illustrates an example of the suggested terminology for different *JK*A* alleles classified based on the sequence patterns.

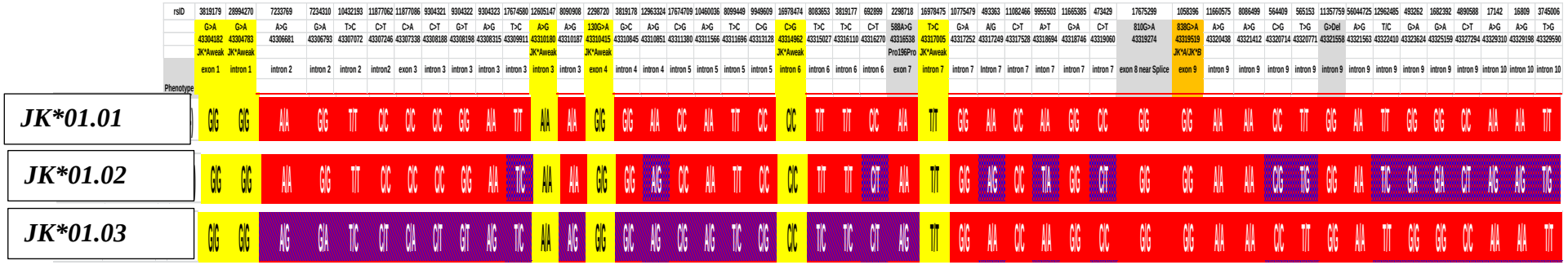


Figure 6.1 Different *JK*A* alleles based on the intronic polymorphism patterns.

Different *JK*A* alleles are described here based on the distinct polymorphism patterns, upon which representative terminology is suggested. For example, *JK*01.01* is suggested for the reference *JK*A* reference sequence whereas *JK*01.02* and *JK*01.03* for the other *JK*A* alleles that differ from the reference. The allele assignment *JK*01.01*,*02* and *03* was based on the frequencies (6 *JK*01.02*, 1 *JK*01.03* and 29 *JK*01.01*) of the *JK*A* haplotype. This figure is a section from (Table 4.5).

6.9 Conclusion

NGS has been shown to be a feasible approach for comprehensively genotyping the blood group genes *FY*, *JK* and *ABO* in addition to other genes, namely *KEL*, *RHD* and *RHCE*, which have been previously sequenced with a similar approach (NGS with LR-PCR) (Avent et al., 2015). NGS-based genotyping allows an accurate reflection and prediction of the phenotype from the genotype by providing the entire encoding molecular sequence responsible for antigen expression for the specialists to analyse. Accordingly, the elucidation of the unusual cases, such as those with hybrid and novel alleles can be feasible. The mass-scale accurate genotyping of donors allows improved blood transfusion safety with better inventory of matched blood units and thereby minimises alloimmunisation and HTR. It is predicted that NGS-based genotyping will replace not only microarray-based genotyping but also serology in the blood group typing of individuals, with great advancements in technology and molecular knowledge being expected in the near future.

References

AGENA BIOSCIENCE. 2017a. *Hemo ID™ Blood Group Genotyping Panel* [Online]. Available: http://agenabio.com/ [Accessed 2017].

AGENA BIOSCIENCE. 2017b. *Hemo ID™ DONOR QUICK SCREEN (DQS) PANEL* [Online]. Available: http://agenabio.com/ [Accessed 2017].

AHN, S. M., KIM, T. H., LEE, S., KIM, D., GHANG, H., KIM, D. S., KIM, B. C., KIM, S. Y., KIM, W. Y., KIM, C., PARK, D., LEE, Y. S., KIM, S., REJA, R., JHO, S., KIM, C. G., CHA, J. Y., KIM, K. H., LEE, B., BHAK, J. & KIM, S. J. 2009. The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res,* 19**,** 1622-9.

AJAY, S. S., PARKER, S. C., ABAAN, H. O., FAJARDO, K. V. & MARGULIES, E. H. 2011. Accurate and comprehensive sequencing of personal genomes. *Genome Res,* 21**,** 1498-505.

ANDREWS, S. 2016. *FastQC: A quality control tool for high throughput sequence data* [Online]. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ [Accessed April 2016].

ANSTEE, D. J. 2009. Red cell genotyping and the future of pretransfusion testing. *Blood,* 114**,** 248-56.

AVENT, N. D. 2007. Large scale blood group genotyping. *Transfus Clin Biol,* 14**,** 10-5.

AVENT, N. D. 2009. Large-scale blood group genotyping: clinical implications. *Br J Haematol,* 144**,** 3-13.

AVENT, N. D., LIU, W., JONES, J. W., SCOTT, M. L., VOAK, D., PISACKA, M., WATT, J. & FLETCHER, A. 1997. Molecular analysis of Rh transcripts and polypeptides from individuals expressing the DVI variant phenotype: an RHD gene deletion event does not generate All DVIccEe phenotypes. *Blood,* 89**,** 1779-86.

AVENT, N. D., MADGETT, T. E., HALAWANI, A. J., ALTAYAR, M. A., KIERNAN, M., REYNOLDS, A. J. & LI, X. 2015. Next-generation sequencing: academic overkill or high-resolution routine blood group genotyping? *ISBT Science Series,* 10**,** 250-256.

AVENT, N. D., MARTINEZ, A., FLEGEL, W. A., OLSSON, M. L., SCOTT, M. L., NOGUÉS, N., PÍSĂCKA, M., DANIELS, G., VAN DER SCHOOT, E.,

MUÑIZ-DIAZ, E., MADGETT, T. E., STORRY, J. R., BEIBOER, S. H., MAASKANT-VAN WIJK, P. A., VON ZABERN, I., JIMÉNEZ, E., TEJEDOR, D., LÓPEZ, M., CAMACHO, E., CHEROUTRE, G., HACKER, A., JINOCH, P., SVOBODOVA, I. & DE HAAS, M. 2007. The BloodGen project: toward mass-scale comprehensive genotyping of blood donors in the European Union and beyond. *Transfusion,* 47**,** 40S-46S.

AVENT, N. D., MARTINEZ, A., FLEGEL, W. A., OLSSON, M. L., SCOTT, M. L., NOGUES, N., PISACKA, M., DANIELS, G. L., MUNIZ-DIAZ, E., MADGETT, T. E., STORRY, J. R., BEIBOER, S., MAASKANT-VAN WIJK, P. M., VON ZABERN, I., JIMENEZ, E., TEJEDOR, D., LOPEZ, M., CAMACHO, E., CHEROUTRE, G., HACKER, A., JINOCH, P., SVOBODOVA, I., VAN DER SCHOOT, E. & DE HAAS, M. 2009. The Bloodgen Project of the European Union, 2003-2009. *Transfus Med Hemother,* 36**,** 162-167.

AVENT, N. D., RIDGWELL, K., TANNER, M. J. & ANSTEE, D. J. 1990. cDNA cloning of a 30 kDa erythrocyte membrane protein associated with Rh (Rhesus)-blood-group-antigen expression. *Biochem J,* 271**,** 821-5.

BACHELERIE, F., BEN-BARUCH, A., BURKHARDT, A. M., COMBADIERE, C., FARBER, J. M., GRAHAM, G. J., HORUK, R., SPARRE-ULRICH, A. H., LOCATI, M., LUSTER, A. D., MANTOVANI, A., MATSUSHIMA, K., MURPHY, P. M., NIBBS, R., NOMIYAMA, H., POWER, C. A., PROUDFOOT, A. E., ROSENKILDE, M. M., ROT, A., SOZZANI, S., THELEN, M., YOSHIE, O. & ZLOTNIK, A. 2014. International Union of Basic and Clinical Pharmacology. [corrected]. LXXXIX. Update on the extended family of chemokine receptors and introducing a new nomenclature for atypical chemokine receptors. *Pharmacol Rev,* 66**,** 1-79.

BACHELERIE, F., GRAHAM, G. J., LOCATI, M., MANTOVANI, A., MURPHY, P. M., NIBBS, R., ROT, A., SOZZANI, S. & THELEN, M. 2015. An atypical addition to the chemokine receptor nomenclature: IUPHAR Review 15. *Br J Pharmacol,* 172**,** 3945-9.

BAKANAY, S. M., OZTURK, A., ILERI, T., INCE, E., YAVASOGLU, S., AKAR, N., UYSAL, Z. & ARSLAN, O. 2013. Blood group genotyping in multi-transfused patients. *Transfus Apher Sci,* 48**,** 257-61.

BALLIF, B. A., HELIAS, V., PEYRARD, T., MENANTEAU, C., SAISON, C., LUCIEN, N., BOURGOUIN, S., LE GALL, M., CARTRON, J. P. & ARNAUD,

L. 2013. Disruption of SMIM1 causes the Vel- blood type. *EMBO Mol Med,* 5**,** 751-61.

BASU, S., KAUR, R. & KAUR, G. 2011. Hemolytic disease of the fetus and newborn: Current trends and perspectives. *Asian J Transfus Sci,* 5**,** 3-7.

BAUMGARTEN, R., VAN GELDER, W., VAN WINTERSHOVEN, J., MAASKANT-VAN WIJK, P. A. & BECKERS, E. A. 2006. Recurrent acute hemolytic transfusion reactions by antibodies against Doa antigens, not detected by cross-matching. *Transfusion,* 46**,** 244-9.

BEIBOER, S. H., WIERINGA-JELSMA, T., MAASKANT-VAN WIJK, P. A., VAN DER SCHOOT, C. E., VAN ZWIETEN, R., ROOS, D., DEN DUNNEN, J. T. & DE HAAS, M. 2005. Rapid genotyping of blood group antigens by multiplex polymerase chain reaction and DNA microarray hybridization. *Transfusion,* 45**,** 667-79.

BENNETT, E. P., STEFFENSEN, R., CLAUSEN, H., WEGHUIS, D. O. & VAN KESSEL, A. G. 1995. Genomic cloning of the human histo-blood group ABO locus. *Biochem Biophys Res Commun,* 206**,** 318-25.

BENTLEY, D. R., BALASUBRAMANIAN, S., SWERDLOW, H. P., SMITH, G. P., MILTON, J., BROWN, C. G., HALL, K. P., EVERS, D. J., BARNES, C. L., BIGNELL, H. R., BOUTELL, J. M., BRYANT, J., CARTER, R. J., KEIRA CHEETHAM, R., COX, A. J., ELLIS, D. J., FLATBUSH, M. R., GORMLEY, N. A., HUMPHRAY, S. J., IRVING, L. J., KARBELASHVILI, M. S., KIRK, S. M., LI, H., LIU, X., MAISINGER, K. S., MURRAY, L. J., OBRADOVIC, B., OST, T., PARKINSON, M. L., PRATT, M. R., RASOLONJATOVO, I. M., REED, M. T., RIGATTI, R., RODIGHIERO, C., ROSS, M. T., SABOT, A., SANKAR, S. V., SCALLY, A., SCHROTH, G. P., SMITH, M. E., SMITH, V. P., SPIRIDOU, A., TORRANCE, P. E., TZONEV, S. S., VERMAAS, E. H., WALTER, K., WU, X., ZHANG, L., ALAM, M. D., ANASTASI, C., ANIEBO, I. C., BAILEY, D. M., BANCARZ, I. R., BANERJEE, S., BARBOUR, S. G., BAYBAYAN, P. A., BENOIT, V. A., BENSON, K. F., BEVIS, C., BLACK, P. J., BOODHUN, A., BRENNAN, J. S., BRIDGHAM, J. A., BROWN, R. C., BROWN, A. A., BUERMANN, D. H., BUNDU, A. A., BURROWS, J. C., CARTER, N. P., CASTILLO, N., CHIARA, E. C. M., CHANG, S., NEIL COOLEY, R., CRAKE, N. R., DADA, O. O., DIAKOUMAKOS, K. D., DOMINGUEZ-FERNANDEZ, B., EARNSHAW, D. J., EGBUJOR, U. C., ELMORE, D. W., ETCHIN, S. S., EWAN, M. R., FEDURCO, M., FRASER, L.

J., FUENTES FAJARDO, K. V., SCOTT FUREY, W., GEORGE, D., GIETZEN, K. J., GODDARD, C. P., GOLDA, G. S., GRANIERI, P. A., GREEN, D. E., GUSTAFSON, D. L., HANSEN, N. F., HARNISH, K., HAUDENSCHILD, C. D., HEYER, N. I., HIMS, M. M., HO, J. T., HORGAN, A. M., et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature,* 456**,** 53-9.

BERGET, S. M. 1995. Exon recognition in vertebrate splicing. *J Biol Chem,* 270**,** 2411-4.

BHUVANESHWAR, K., SULAKHE, D., GAUBA, R., RODRIGUEZ, A., MADDURI, R., DAVE, U., LACINSKI, L., FOSTER, I., GUSEV, Y. & MADHAVAN, S. 2015. A case study for cloud based high throughput analysis of NGS data using the globus genomics system. *Comput Struct Biotechnol J,* 13**,** 64-74.

BRAGG, L. M., STONE, G., BUTLER, M. K., HUGENHOLTZ, P. & TYSON, G. W. 2013. Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. *PLoS Comput Biol,* 9**,** e1003031.

BUGERT, P., MCBRIDE, S., SMITH, G., DUGRILLON, A., KLUTER, H., OUWEHAND, W. H. & METCALFE, P. 2005. Microarray-based genotyping for blood groups: comparison of gene array and 5'-nuclease assay techniques with human platelet antigen as a model. *Transfusion,* 45**,** 654-9.

CARTRON, J. P., BAILLY, P., LE VAN KIM, C., CHERIF-ZAHAR, B., MATASSI, G., BERTRAND, O. & COLIN, Y. 1998. Insights into the structure and function of membrane polypeptides carrying blood group antigens. *Vox Sang,* 2**,** 29-64.

CASALI., N. & PRESTON, A. 2003. *E. coli Plasmid Vectors*, Humana Press.

CASAS, J., FRIEDMAN, D. F., JACKSON, T., VEGE, S., WESTHOFF, C. M. & CHOU, S. T. 2015. Changing practice: red blood cell typing by molecular methods for patients with sickle cell disease. *Transfusion,* 55**,** 1388-93.

CASTILHO, L. 2007. The value of DNA analysis for antigens in the Duffy blood group system. *Transfusion,* 47**,** 28S-31S.

CASTILHO, L., RIOS, M., BIANCO, C., PELLEGRINO, J., JR., ALBERTO, F. L., SAAD, S. T. & COSTA, F. F. 2002. DNA-based typing of blood groups for the management of multiply-transfused sickle cell disease patients. *Transfusion,* 42**,** 232-8.

CHAUDHURI, A., NIELSEN, S., ELKJAER, M. L., ZBRZEZNA, V., FANG, F. & POGO, A. O. 1997. Detection of Duffy antigen in the plasma membranes and caveolae of vascular endothelial and epithelial cells of nonerythroid organs.

Blood, 89**,** 701-12.

CHAUDHURI, A., POLYAKOVA, J., ZBRZEZNA, V. & POGO, A. O. 1995. The coding sequence of Duffy blood group gene in humans and simians: restriction fragment length polymorphism, antibody and malarial parasite specificities, and expression in nonerythroid tissues in Duffy-negative individuals. *Blood,* 85**,** 615-21.

CHAUDHURI, A., POLYAKOVA, J., ZBRZEZNA, V., WILLIAMS, K., GULATI, S. & POGO, A. O. 1993. Cloning of glycoprotein D cDNA, which encodes the major subunit of the Duffy blood group system and the receptor for the Plasmodium vivax malaria parasite. *Proc Natl Acad Sci U S A,* 90**,** 10793-7.

CHESTER, M. A. & OLSSON, M. L. 2001. The ABO blood group gene: a locus of considerable genetic diversity. *Transfus Med Rev,* 15**,** 177-200.

CHOU, S. T. & WESTHOFF, C. M. 2011. The role of molecular immunohematology in sickle cell disease. *Transfus Apher Sci,* 44**,** 73-9.

CLAUSEN, H., BENNETT, E. P. & GRUNNET, N. 1994. Molecular genetics of ABO histo-blood groups. *Transfus Clin Biol,* 1**,** 79-89.

CONSORTIUM, I. H. G. 2004. Finishing the euchromatic sequence of the human genome. *Nature,* 431**,** 931-45.

COOMBS, R. R. & ROBERTS, F. 1959. The antiglobulin reaction. *Br Med Bull,* 15**,** 113-8.

COOPER, D. N. & GROUP, N. P. 2003. *Nature Encyclopedia of the Human Genome: Genome databases - Mitochondrial genome: Evolution,* Nature Publishing Group.

CUTBUSH, M. & MOLLISON, P. L. 1950. The Duffy blood group system. *Heredity,* 4**,** 383-9.

CVEJIC, A., HAER-WIGMAN, L., STEPHENS, J. C., KOSTADIMA, M., SMETHURST, P. A., FRONTINI, M., VAN DEN AKKER, E., BERTONE, P., BIELCZYK-MACZYNSKA, E., FARROW, S., FEHRMANN, R. S., GRAY, A., DE HAAS, M., HAVER, V. G., JORDAN, G., KARJALAINEN, J., KERSTENS, H. H., KIDDLE, G., LLOYD-JONES, H., NEEDS, M., POOLE, J., SOUSSAN, A. A., RENDON, A., RIENECK, K., SAMBROOK, J. G., SCHEPERS, H., SILLJE, H. H., SIPOS, B., SWINKELS, D., TAMURI, A. U., VERWEIJ, N., WATKINS, N. A., WESTRA, H. J., STEMPLE, D., FRANKE, L., SORANZO, N., STUNNENBERG, H. G., GOLDMAN, N., VAN DER HARST, P., VAN DER SCHOOT, C. E., OUWEHAND, W. H. & ALBERS, C.

A. 2013. SMIM1 underlies the Vel blood group and influences red blood cell traits. *Nat Genet,* 45**,** 542-5.

DABELSTEEN, E. & GAO, S. 2005. ABO blood-group antigens in oral cancer. *J Dent Res,* 84**,** 21-8.

DANIELS, G. 2005. The molecular genetics of blood group polymorphism. *Transpl Immunol,* 14**,** 143-53.

DANIELS, G. 2013. Human Blood Groups: Introduction. *Human Blood Groups.* Wiley-Blackwell.

DANIELS, G., FINNING, K., MARTIN, P. & MASSEY, E. 2009. Noninvasive prenatal diagnosis of fetal blood group phenotypes: current practice and future prospects. *Prenat Diagn,* 29**,** 101-7.

DANIELS, G., FLEGEL, W. A., FLETCHER, A., GARRATTY, G., LEVENE, C., LOMAS-FRANCIS, C., MOULDS, J. M., MOULDS, J. J., OLSSON, M. L., OVERBEEKE, M. A., POOLE, J., REID, M. E., ROUGER, P., VAN DER SCHOOT, C. E., SCOTT, M., SISTONEN, P., SMART, E., STORRY, J. R., TANI, Y., YU, L. C., WENDEL, S., WESTHOFF, C. M. & ZELINSKI, T. 2007. International Society of Blood Transfusion Committee on Terminology for Red Cell Surface Antigens: Cape Town report. *Vox Sang,* 92**,** 250-3.

DANIELS, G. L., FLETCHER, A., GARRATTY, G., HENRY, S., JORGENSEN, J., JUDD, W. J., LEVENE, C., LOMAS-FRANCIS, C., MOULDS, J. J., MOULDS, J. M., MOULDS, M., OVERBEEKE, M., REID, M. E., ROUGER, P., SCOTT, M., SISTONEN, P., SMART, E., TANI, Y., WENDEL, S. & ZELINSKI, T. 2004. Blood group terminology 2004: from the International Society of Blood Transfusion committee on terminology for red cell surface antigens. *Vox Sang,* 87**,** 304-16.

DAVENPORT, R. D. 2009. An introduction to chemokines and their roles in transfusion medicine. *Vox Sang,* 96**,** 183-98.

DBRBC. 2016. *Blood Group Antigen Gene Mutation Database (BGMUT)* [Online]. Available: http://www.ncbi.nlm.nih.gov/projects/gv/mhc/xslcgi.cgi?cmd=bgmut/home [Accessed 2016].

DEAN, L. 2005. *Blood Groups and Red Cell Antigens [Internet]*, National Center for Biotechnology Information (NCBI).

DONAHUE, R. P., BIAS, W. B., RENWICK, J. H. & MCKUSICK, V. A. 1968. Probable assignment of the Duffy blood group locus to chromosome 1 in man.

Proc Natl Acad Sci U S A, 61**,** 949-55.

EWING, B., HILLIER, L., WENDL, M. C. & GREEN, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res,* 8**,** 175-85.

FERRANDO, M., MARTINEZ-CANABATE, S., LUNA, I., DE LA RUBIA, J., CARPIO, N., ALFREDO, P. & ARRIAGA, F. 2008. Severe hemolytic disease of the fetus due to anti-Jkb. *Transfusion,* 48**,** 402-4.

FICHOU, Y., AUDREZET, M. P., GUEGUEN, P., LE MARECHAL, C. & FEREC, C. 2014. Next-generation sequencing is a credible strategy for blood group genotyping. *Br J Haematol,* 167**,** 554-62.

FINNING, K., MARTIN, P., SUMMERS, J., MASSEY, E., POOLE, G. & DANIELS, G. 2008. Effect of high throughput RHD typing of fetal DNA in maternal plasma on use of anti-RhD immunoglobulin in RhD negative pregnant women: prospective feasibility study. *BMJ,* 336**,** 816-8.

FLEGEL, W. A., GOTTSCHALL, J. L. & DENOMME, G. A. 2015. Implementing mass-scale red cell genotyping at a blood center. *Transfusion,* 55**,** 2610-5; quiz 2609.

FUKUMORI, Y., OHNOKI, S., SHIBATA, H., YAMAGUCHI, H. & NISHIMUKAI, H. 1995. Genotyping of ABO blood groups by PCR and RFLP analysis of 5 nucleotide positions. *Int J Legal Med,* 107**,** 179-82.

GASSNER, C., KRAUS, R. L., DOVC, T., KILGA-NOGLER, S., UTZ, I., MUELLER, T. H., SCHUNTER, F. & SCHOENITZER, D. 2000. Fyx is associated with two missense point mutations in its gene and can be detected by PCR-SSP. *Immunohematology,* 16**,** 61-7.

GENOMES PROJECT, C., ABECASIS, G. R., ALTSHULER, D., AUTON, A., BROOKS, L. D., DURBIN, R. M., GIBBS, R. A., HURLES, M. E. & MCVEAN, G. A. 2010. A map of human genome variation from population-scale sequencing. *Nature,* 467**,** 1061-73.

GIANGRANDE, P. L. 2000. The history of blood transfusion. *Br J Haematol,* 110**,** 758-67.

GIOLLO, M., MINERVINI, G., SCALZOTTO, M., LEONARDI, E., FERRARI, C. & TOSATTO, S. C. 2015. BOOGIE: Predicting Blood Groups from High Throughput Sequencing Data. *PLoS One,* 10**,** e0124579.

GOODRICK, M. J., HADLEY, A. G. & POOLE, G. 1997. Haemolytic disease of the fetus and newborn due to anti-Fya and the potential clinical value of Duffy

genotyping in pregnancies at risk. *Transfusion Medicine,* 7**,** 301-304.

GOODWIN, S., MCPHERSON, J. D. & MCCOMBIE, W. R. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet,* 17**,** 333-51.

GREEN, C. 1989. The ABO, Lewis and related blood group antigens; a review of structure and biosynthesis. *FEMS Microbiol Immunol,* 1**,** 321-30.

HADLEY, T. J., LU, Z. H., WASNIOWSKA, K., MARTIN, A. W., PEIPER, S. C., HESSELGESSER, J. & HORUK, R. 1994. Postcapillary venule endothelial cells in kidney express a multispecific chemokine receptor that is structurally and functionally identical to the erythroid isoform, which is the Duffy blood group antigen. *J Clin Invest,* 94**,** 985-91.

HALAWANI, A. J. 2015. *THE FUTURE OF NEXT-GENERATION SEQUENCING FOR BLOOD GROUP GENOTYPING AND ITS IMPLICATIONS IN TRANSFUSION MEDICINE.* PhD, Plymouth University.

HASHMI, G., SHARIFF, T., SEUL, M., VISSAVAJJHALA, P., HUE-ROYE, K., CHARLES-PIERRE, D., LOMAS-FRANCIS, C., CHAUDHURI, A. & REID, M. E. 2005. A flexible array format for large-scale, rapid blood group DNA typing. *Transfusion,* 45**,** 680-8.

HASHMI, G., SHARIFF, T., ZHANG, Y., CRISTOBAL, J., CHAU, C., SEUL, M., VISSAVAJJHALA, P., BALDWIN, C., HUE-ROYE, K., CHARLES-PIERRE, D., LOMAS-FRANCIS, C. & REID, M. E. 2007. Determination of 24 minor red blood cell antigens for more than 2000 blood donors by high-throughput DNA analysis. *Transfusion,* 47**,** 736-47.

HEATON, D. C. & MCLOUGHLIN, K. 1982. Jk(a-b-) red blood cells resist urea lysis. *Transfusion,* 22**,** 70-1.

HELIAS, V., SAISON, C., BALLIF, B. A., PEYRARD, T., TAKAHASHI, J., TAKAHASHI, H., TANAKA, M., DEYBACH, J. C., PUY, H., LE GALL, M., SUREAU, C., PHAM, B. N., LE PENNEC, P. Y., TANI, Y., CARTRON, J. P. & ARNAUD, L. 2012. ABCB6 is dispensable for erythropoiesis and specifies the new blood group system Langereis. *Nat Genet,* 44**,** 170-3.

HENNY, C., LEJON CROTTET, S., GOWLAND, P. L. & NIDERHAUSER, C. H., H. 2014. Three novel JK alleles detected in Swiss blood donors. *Vox Sang,* 107, 188.

HGNC. 2016. *HUGO Gene Nomenclature Committee* [Online]. Available: http://www.genenames.org/ [Accessed 2016].

HODKINSON, B. P. & GRICE, E. A. 2015. Next-Generation Sequencing: A Review of

Technologies and Tools for Wound Microbiome Research. *Adv Wound Care (New Rochelle),* 4**,** 50-58.

HORN, T., CASTILHO, L., MOULDS, J. M., BILLINGSLEY, K., VEGE, S., JOHNSON, N. & WESTHOFF, C. M. 2012. A novel JKA allele, nt561C>A, associated with silencing of Kidd expression. *Transfusion,* 52**,** 1092-6.

HOSOI, E. 2008. Biological and clinical aspects of ABO blood group system. *J Med Invest,* 55**,** 174-82.

HOSSEINI-MAAF, B., HELLBERG, A., RODRIGUES, M. J., CHESTER, M. A. & OLSSON, M. L. 2003. ABO exon and intron analysis in individuals with the AweakB phenotype reveals a novel O1v-A2 hybrid allele that causes four missense mutations in the A transferase. *BMC Genet,* 4**,** 17.

HOSSEINI-MAAF, B., IRSHAID, N. M., HELLBERG, A., WAGNER, T., LEVENE, C., HUSTINX, H., STEFFENSEN, R., CHESTER, M. A. & OLSSON, M. L. 2005. New and unusual O alleles at the ABO locus are implicated in unexpected blood group phenotypes. *Transfusion,* 45**,** 70-81.

HTTP://BLAST.NCBI.NLM.NIH.GOV/.

HTTP://IONCOMMUNITY.LIFETECHNOLOGIES.COM/MESSAGE/15695#15695.

HUH, J. Y., PARK, G., JANG, S. J., MOON, D. S. & PARK, Y. J. 2011. A rapid long PCR-direct sequencing analysis for ABO genotyping. *Ann Clin Lab Sci,* 41**,** 340-5.

HURD, P. J. & NELSON, C. J. 2009. Advantages of next-generation sequencing versus the microarray in epigenetic research. *Brief Funct Genomic Proteomic,* 8**,** 174-83.

HUSSAIN, S. S., EBBS, A. M., CURTIN, N. J. & KEIDAN, A. J. 2007. Delayed haemolytic transfusion reaction due to anti-Jkb in a patient with non-Hodgkin's lymphoma-transient nature of anti-Jkb and the importance of early serological diagnosis. *Transfus Med,* 17**,** 197-9.

IKIN, E. W., PRIOR, A. M., RACE, R. R. & TAYLOR, G. L. 1939. THE DISTRIBUTIONS IN THE A1A2BO BLOOD GROUPS IN ENGLAND. *Annals of Eugenics,* 9**,** 409-411.

INNO-TRAIN. 2017. *RBC-FLUOGENE - THE FLUORESCENCE PCR* [Online]. Available: http://www.inno-train.de/en/products/bloodgroup-detection/rbc-fluogene [Accessed 2017].

IRSHAID, N. M., CHESTER, M. A. & OLSSON, M. L. 1999. Allele-related variation in minisatellite repeats involved in the transcription of the blood group ABO

gene. *Transfus Med,* 9**,** 219-26.

IRSHAID, N. M., EICHER, N. I., HUSTINX, H., POOLE, J. & OLSSON, M. L. 2002. Novel alleles at the JK blood group locus explain the absence of the erythrocyte urea transporter in European families. *Br J Haematol,* 116**,** 445-53.

IRSHAID, N. M., HENRY, S. M. & OLSSON, M. L. 2000. Genomic characterization of the Kidd blood group gene:different molecular basis of the Jk(a–b–) phenotypein Polynesians and Finns. *Transfusion,* 40**,** 69-74.

ISBT. 2016. *Red Cell Immunogenetics and Blood Group Terminology* [Online]. Available: http://www.isbtweb.org/ [Accessed August 2016].

IWAMOTO, S., LI, J., OMI, T., IKEMOTO, S. & KAJII, E. 1996a. Identification of a novel exon and spliced form of Duffy mRNA that is the predominant transcript in both erythroid and postcapillary venule endothelium. *Blood,* 87**,** 378-385.

IWAMOTO, S., LI, J., SUGIMOTO, N., OKUDA, H. & KAJII, E. 1996b. Characterization of the Duffy Gene Promotor: Evidence for Tissue-Specific Abolishment of Expression in Fy(a−b−) of Black Individuals. *Biochemical and Biophysical Research Communications,* 222**,** 852-859.

IWAMOTO, S., OMI, T., KAJII, E. & IKEMOTO, S. 1995. Genomic organization of the glycoprotein D gene: Duffy blood group Fya/Fyb alloantigen system is associated with a polymorphism at the 44-amino acid residue. *Blood,* 85**,** 622-6.

JANATPOUR, K. A., KALMIN, N. D., JENSEN, H. M. & HOLLAND, P. V. 2008. Clinical outcomes of ABO-incompatible RBC transfusions. *Am J Clin Pathol,* 129**,** 276-81.

JUNGBAUER, C., HOBEL, C. M., SCHWARTZ, D. W. & MAYR, W. R. 2012. High-throughput multiplex PCR genotyping for 35 red blood cell antigens in blood donors. *Vox Sang,* 102**,** 234-42.

KELLER, M., CROWLER JA & T, H. 2014. Kidd antigen discrepancies: genotype-predicted phenotype vs serological phenotype. *Vox Sang,* 107 (S1):37.

KING, C. L., MICHON, P., SHAKRI, A. R., MARCOTTY, A., STANISIC, D., ZIMMERMAN, P. A., COLE-TOBIAN, J. L., MUELLER, I. & CHITNIS, C. E. 2008. Naturally acquired Duffy-binding protein-specific binding inhibitory antibodies confer protection from blood-stage Plasmodium vivax infection. *Proc Natl Acad Sci U S A,* 105**,** 8363-8.

KOMINATO, Y., TSUCHIYA, T., HATA, N., TAKIZAWA, H. & YAMAMOTO, F. 1997. Transcription of human ABO histo-blood group genes is dependent upon binding of transcription factor CBF/NF-Y to minisatellite sequence. *J Biol Chem,*

272**,** 25890-8.

KUMPEL, B. M. 2008. Lessons learnt from many years of experience using anti-D in humans for prevention of RhD immunization and haemolytic disease of the fetus and newborn. *Clin Exp Immunol,* 154**,** 1-5.

LANDSTEINER, K. 1961. On agglutination of normal human blood. *Transfusion,* 1**,** 5-8.

LANE, W. J., WESTHOFF, C. M., UY, J. M., AGUAD, M., SMELAND-WAGMAN, R., KAUFMAN, R. M., REHM, H. L., GREEN, R. C., SILBERSTEIN, L. E. & MEDSEQ, P. 2016. Comprehensive red blood cell and platelet antigen prediction from whole genome sequencing: proof of principle. *Transfusion,* 56**,** 743-54.

LANG, K., WAGNER, I., SCHONE, B., SCHOFL, G., BIRKNER, K., HOFMANN, J. A., SAUTER, J., PINGEL, J., BOHME, I., SCHMIDT, A. H. & LANGE, V. 2016. ABO allele-level frequency estimation based on population-scale genotyping by next generation sequencing. *BMC Genomics,* 17**,** 374.

LANGE, V., BOHME, I., HOFMANN, J., LANG, K., SAUTER, J., SCHONE, B., PAUL, P., ALBRECHT, V., ANDREAS, J. M., BAIER, D. M., NETHING, J., EHNINGER, U., SCHWARZELT, C., PINGEL, J., EHNINGER, G. & SCHMIDT, A. H. 2014. Cost-efficient high-throughput HLA typing by MiSeq amplicon sequencing. *BMC Genomics,* 15**,** 63.

LE VAN KIM, C., MOURO, I., CHERIF-ZAHAR, B., RAYNAL, V., CHERRIER, C., CARTRON, J. P. & COLIN, Y. 1992. Molecular cloning and primary structure of the human blood group RhD polypeptide. *Proc Natl Acad Sci U S A,* 89**,** 10925-9.

LEE, S., RUSSO, D. C., REINER, A. P., LEE, J. H., SY, M. Y., TELEN, M. J., JUDD, W. J., SIMON, P., RODRIGUES, M. J., CHABERT, T., POOLE, J., JOVANOVIC-SRZENTIC, S., LEVENE, C., YAHALOM, V. & REDMAN, C. M. 2001. Molecular defects underlying the Kell null phenotype. *J Biol Chem,* 276**,** 27281-9.

LERUT, E., VAN DAMME, B., NOIZAT-PIRENNE, F., EMONDS, M. P., ROUGER, P., VANRENTERGHEM, Y., PIRENNE, J. & ANSART-PIRENNE, H. 2007. Duffy and Kidd blood group antigens: minor histocompatibility antigens involved in renal allograft rejection? *Transfusion,* 47**,** 28-40.

LEVINE, P. 1961. A review of Landsteiner's contributions to human blood groups. *Transfusion,* 1**,** 45-52.

LIFERECHNOLOGIES. 2012. *PGM™ for genes. Proton™ for genomes* [Online]. Available: https://tools.thermofisher.com/content/sfs/brochures/PGM-Proton-sequencers.pdf [Accessed].

LIU, L., LI, Y., LI, S., HU, N., HE, Y., PONG, R., LIN, D., LU, L. & LAW, M. 2012. Comparison of next-generation sequencing systems. *J Biomed Biotechnol,* 2012**,** 251364.

LOMAN, N. J., MISRA, R. V., DALLMAN, T. J., CONSTANTINIDOU, C., GHARBIA, S. E., WAIN, J. & PALLEN, M. J. 2012. Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol,* 30**,** 434-9.

LOMAS-FRANCIS, C. 2007. The value of DNA analysis for antigens of the Kidd blood group system. *Transfusion,* 47**,** 23S-7S.

LUCIEN, N., CHIARONI, J., CARTRON, J. P. & BAILLY, P. 2002a. Partial deletion in the JK locus causing a Jk(null) phenotype. *Blood,* 99**,** 1079-81.

LUCIEN, N., SIDOUX-WALTER, F., OLIVÈS, B., MOULDS, J., LE PENNEC, P.-Y., CARTRON, J.-P. & BAILLY, P. 1998. Characterization of the Gene Encoding the Human Kidd Blood Group/Urea Transporter Protein: EVIDENCE FOR SPLICE SITE MUTATIONS IN JknullINDIVIDUALS. *Journal of Biological Chemistry,* 273**,** 12973-12980.

LUCIEN, N., SIDOUX-WALTER, F., ROUDIER, N., RIPOCHE, P., HUET, M., TRINH-TRANG-TAN, M.-M., CARTRON, J.-P. & BAILLY, P. 2002b. Antigenic and Functional Properties of the Human Red Blood Cell Urea Transporter hUT-B1. *Journal of Biological Chemistry,* 277**,** 34101-34108.

MALLINSON, G., SOO, K. S., SCHALL, T. J., PISACKA, M. & ANSTEE, D. J. 1995. Mutations in the erythrocyte chemokine receptor (Duffy) gene: the molecular basis of the Fya/Fyb antigens and identification of a deletion in the Duffy gene of an apparently healthy individual with the Fy(a-b-) phenotype. *Br J Haematol,* 90**,** 823-9.

MARDIS, E. R. 2013. Next-generation sequencing platforms. *Annu Rev Anal Chem (Palo Alto Calif),* 6**,** 287-303.

MARGULIES, M., EGHOLM, M., ALTMAN, W. E., ATTIYA, S., BADER, J. S., BEMBEN, L. A., BERKA, J., BRAVERMAN, M. S., CHEN, Y. J., CHEN, Z., DEWELL, S. B., DU, L., FIERRO, J. M., GOMES, X. V., GODWIN, B. C., HE, W., HELGESEN, S., HO, C. H., IRZYK, G. P., JANDO, S. C., ALENQUER, M. L., JARVIE, T. P., JIRAGE, K. B., KIM, J. B., KNIGHT, J. R., LANZA, J. R., LEAMON, J. H., LEFKOWITZ, S. M., LEI, M., LI, J., LOHMAN, K. L., LU,

H., MAKHIJANI, V. B., MCDADE, K. E., MCKENNA, M. P., MYERS, E. W., NICKERSON, E., NOBILE, J. R., PLANT, R., PUC, B. P., RONAN, M. T., ROTH, G. T., SARKIS, G. J., SIMONS, J. F., SIMPSON, J. W., SRINIVASAN, M., TARTARO, K. R., TOMASZ, A., VOGT, K. A., VOLKMER, G. A., WANG, S. H., WANG, Y., WEINER, M. P., YU, P., BEGLEY, R. F. & ROTHBERG, J. M. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature,* 437**,** 376-80.

MARSHALL, C. S., DWYRE, D., ECKERT, R. & RUSSELL, L. 1999. Severe hemolytic reaction due to anti-JK3. *Arch Pathol Lab Med,* 123**,** 949-51.

MAYR, F. B., SPIEL, A. O., LEITNER, J. M., FIRBAS, C., KLIEGEL, T., JILMA-STOHLAWETZ, P., DERENDORF, H. & JILMA, B. 2008. Duffy antigen modifies the chemokine response in human endotoxemia. *Crit Care Med,* 36**,** 159-65.

MAZONSON, P., EFRUSY, M., SANTAS, C., ZIMAN, A., BURNER, J., ROSEFF, S., VIJAYARAGHAVAN, A. & KAUFMAN, R. 2014. The HI-STAR study: resource utilization and costs associated with serologic testing for antibody-positive patients at four United States medical centers. *Transfusion,* 54**,** 271-7.

MCBEAN, R. S., HYLAND, C. A. & FLOWER, R. L. 2014. Approaches to determination of a full profile of blood group genotypes: single nucleotide variant mapping and massively parallel sequencing. *Comput Struct Biotechnol J,* 11**,** 147-51.

MENY, G. M. 2010. The Duffy blood group system: a review. *Immunohematology,* 26**,** 51-6.

MERRIMAN, B., ION TORRENT, R., TEAM, D. & ROTHBERG, J. M. 2012. Progress in ion torrent semiconductor chip based sequencing. *Electrophoresis,* 33**,** 3397-417.

METZKER, M. L. 2010. Sequencing technologies - the next generation. *Nat Rev Genet,* 11**,** 31-46.

MILLER, L. H., MASON, S. J., CLYDE, D. F. & MCGINNISS, M. H. 1976. The Resistance Factor to Plasmodium vivax in Blacks. *New England Journal of Medicine,* 295**,** 302-304.

MILLER, L. H., MASON, S. J., DVORAK, J. A., MCGINNISS, M. H. & ROTHMAN, I. K. 1975. Erythrocyte receptors for (Plasmodium knowlesi) malaria: Duffy blood group determinants. *Science,* 189**,** 561-3.

MOLLICONE, R., CAILLEAU, A. & ORIOL, R. 1995. Molecular genetics of H, Se,

Lewis and other fucosyltransferase genes. *Transfus Clin Biol,* 2**,** 235-42.

MOORES, P. P., ISSITT, P. D., PAVONE, B. G. & MCKEEVER, B. G. 1975. Some observations on "Bombay" bloods, with comments on evidence for the existence of two different Oh phenotypes. *Transfusion,* 15**,** 237-43.

MORGAN, W. T. & WATKINS, W. M. 1948. The detection of a product of the blood group O gene and the relationship of the so-called O-substance to the agglutinates A and B. *Br J Exp Pathol,* 29**,** 159-73.

MULLINS, F. M., DIETZ, L., LAY, M., ZEHNDER, J. L., FORD, J., CHUN, N. & SCHRIJVER, I. 2007. Identification of an intronic single nucleotide polymorphism leading to allele dropout during validation of a CDH1 sequencing assay: implications for designing polymerase chain reaction-based assays. *Genet Med,* 9**,** 752-760.

MURPHY, M. T., TEMPLETON, L. J., FLEMING, J., FERGUSON, M., PETERKIN, M. & FRASER, R. H. 1997. Comparison of Fyb status as determined serologically and genetically. *Transfusion Medicine,* 7**,** 135-141.

NCBI. 2016a. *dbSNP* [Online]. Available: http://www.ncbi.nlm.nih.gov/SNP/ [Accessed 2016].

NCBI. 2016b. *http://www.ncbi.nlm.nih.gov/gene* [Online]. Available: http://www.ncbi.nlm.nih.gov/gene [Accessed 2016].

NCBI. 2016c. *National Center for Biotechnology Information* [Online]. Available: http://www.ncbi.nlm.nih.gov/ [Accessed 2013-2016].

NHGRI. 2016. *National Human Genome Research Institute* [Online]. Available: www.genome.gov [Accessed 2016].

OLIVÈS, B., MARTIAL, S., MATTEI, M.-G., MATASSI, G., ROUSSELET, G., RIPOCHE, P., CARTRON, J.-P. & BAILLY, P. 1996. Molecular characterization of a new urea transporter in the human kidney. *FEBS Letters,* 386**,** 156-160.

OLIVÈS, B., MERRIMAN, M., BAILLY, P., BAIN, S., BARNETT, A., TODD, J., CARTRON, J.-P. & MERRIMAN, T. 1997. The Molecular Basis of the Kidd Blood Group Polymorphism and Its Lack of Association With Type 1 Diabetes Susceptibility. *Human Molecular Genetics,* 6**,** 1017-1020.

OLSSON, M. L. & CHESTER, M. A. 1996. Frequent occurrence of a variant O1 gene at the blood group ABO locus. *Vox Sang,* 70**,** 26-30.

OLSSON, M. L. & CHESTER, M. A. 1998. Heterogeneity of the blood group Ax allele: genetic recombination of common alleles can result in the Ax phenotype.

Transfus Med, 8**,** 231-8.

OLSSON, M. L., IRSHAID, N. M., HOSSEINI-MAAF, B., HELLBERG, A., MOULDS, M. K., SARENEVA, H. & CHESTER, M. A. 2001. Genomic analysis of clinical samples with serologic ABO blood grouping discrepancies: identification of 15 novel A and B subgroup alleles. *Blood,* 98**,** 1585-93.

OLSSON, M. L., SMYTHE, J. S., HANSSON, C., POOLE, J., MALLINSON, G., JONES, J., AVENT, N. D. & DANIELS, G. 1998. The Fyx phenotype is associated with a missense mutation in the Fyb allele predicting Arg89Cys in the Duffy glycoprotein. *Br J Haematol,* 103**,** 1184-1191.

ONSONGO, G., ERDMANN, J., SPEARS, M. D., CHILTON, J., BECKMAN, K. B., HAUGE, A., YOHE, S., SCHOMAKER, M., BOWER, M., SILVERSTEIN, K. A. & THYAGARAJAN, B. 2014. Implementation of Cloud based next generation sequencing data analysis in a clinical laboratory. *BMC Res Notes,* 7**,** 314.

OWEN, R. 2000. Karl Landsteiner and the first human marker locus. *Genetics,* 155**,** 995-8.

PAMPHILON, D. H. & SCOTT, M. L. 2007. Robin Coombs: his life and contribution to haematology and transfusion medicine. *Br J Haematol,* 137**,** 401-8.

PARIS, S., RIGAL, D., BARLET, V., VERDIER, M., COUDURIER, N., BAILLY, P. & BRES, J. C. 2014. Flexible automated platform for blood group genotyping on DNA microarrays. *J Mol Diagn,* 16**,** 335-42.

PATENAUDE, S. I., SETO, N. O., BORISOVA, S. N., SZPACENKO, A., MARCUS, S. L., PALCIC, M. M. & EVANS, S. V. 2002. The structural basis for specificity in human ABO(H) blood group biosynthesis. *Nat Struct Biol,* 9**,** 685-90.

PATNAIK, S. K., HELMBERG, W. & BLUMENFELD, O. O. 2012. BGMUT: NCBI dbRBC database of allelic variations of genes encoding antigens of blood group systems. *Nucleic Acids Res,* 40**,** D1023-9.

PAULSON, J. C. & COLLEY, K. J. 1989. Glycosyltransferases. Structure, localization, and control of cell type-specific glycosylation. *J Biol Chem,* 264**,** 17615-8.

PEIPER, S. C., WANG, Z. X., NEOTE, K., MARTIN, A. W., SHOWELL, H. J., CONKLYN, M. J., OGBORNE, K., HADLEY, T. J., LU, Z. H., HESSELGESSER, J. & HORUK, R. 1995. The Duffy antigen/receptor for chemokines (DARC) is expressed in endothelial cells of Duffy negative individuals who lack the erythrocyte receptor. *J Exp Med,* 181**,** 1311-7.

PERREAULT, J., LAVOIE, J., PAINCHAUD, P., COTE, M., CONSTANZO-YANEZ, J., COTE, R., DELAGE, G., GENDRON, F., DUBUC, S., CARON, B., LEMIEUX, R. & ST-LOUIS, M. 2009. Set-up and routine use of a database of 10,555 genotyped blood donors to facilitate the screening of compatible blood components for alloimmunized patients. *Vox Sang,* 97**,** 61-8.

PINEDA, A. A., VAMVAKAS, E. C., GORDEN, L. D., WINTERS, J. L. & MOORE, S. B. 1999. Trends in the incidence of delayed hemolytic and delayed serologic transfusion reactions. *Transfusion,* 39**,** 1097-103.

PLAUT, G., IKIN, E. W., MOURANT, A. E., SANGER, R. & RACE, R. R. 1953. A new blood-group antibody, anti Jkb. *Nature,* 171**,** 431.

POLIN, H., DANZER, M., PROLL, J., HOFER, K., HEILINGER, U., ZOPF, A. & GABRIEL, C. 2008. Introduction of a real-time-based blood-group genotyping approach. *Vox Sang,* 95**,** 125-30.

POOLE, J. & DANIELS, G. 2007. Blood group antibodies and their significance in transfusion medicine. *Transfus Med Rev,* 21**,** 58-71.

PROBER, J. M., TRAINOR, G. L., DAM, R. J., HOBBS, F. W., ROBERTSON, C. W., ZAGURSKY, R. J., COCUZZA, A. J., JENSEN, M. A. & BAUMEISTER, K. 1987. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science,* 238**,** 336-41.

PROCTER, J., CRAWFORD, J., BUNCE, M. & WELSH, K. I. 1997. A rapid molecular method (polymerase chain reaction with sequence-specific primers) to genotype for ABO blood group and secretor status and its potential for organ transplants. *Tissue Antigens,* 50**,** 475-83.

PROMEGA. 2016. *Gene cloning* [Online]. Promega. Available: https://www.promega.co.uk/resources/product-guides-and-selectors/cloning-enzymes-guide/ [Accessed 01 Sep 2016 2016].

PRUENSTER, M. & ROT, A. 2006. Throwing light on DARC. *Biochem Soc Trans,* 34**,** 1005-8.

QUAIL, M. A., SMITH, M., COUPLAND, P., OTTO, T. D., HARRIS, S. R., CONNOR, T. R., BERTONI, A., SWERDLOW, H. P. & GU, Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics,* 13**,** 341.

REID, M. E. 2009. Transfusion in the age of molecular diagnostics. *Hematology Am Soc Hematol Educ Program***,** 171-7.

REID, M. E. & DENOMME, G. A. 2011. DNA-based methods in the

immunohematology reference laboratory. *Transfus Apher Sci,* 44**,** 65-72.

REID, M. E., LOMAS-FRANCIS, C. & OLSSON, M. L. 2012. 1 - Introduction. *In:* REID, M. E., LOMAS-FRANCIS, C. & OLSSON, M. L. (eds.) *The Blood Group Antigen FactsBook (Third Edition).* Boston: Academic Press.

REID, M. E. & MOHANDAS, N. 2004. Red blood cell blood group antigens: structure and function. *Seminars in Hematology,* 41**,** 93-117.

REID, M. E., RIOS, M., POWELL, V. I., CHARLES-PIERRE, D. & MALAVADE, V. 2000. DNA from blood samples can be used to genotype patients who have recently received a transfusion. *Transfusion,* 40**,** 48-53.

RIENECK, K., BAK, M., JONSON, L., CLAUSEN, F. B., KROG, G. R., TOMMERUP, N., NIELSEN, L. K., HEDEGAARD, M. & DZIEGIEL, M. H. 2013. Next-generation sequencing: proof of concept for antenatal prediction of the fetal Kell blood group phenotype from cell-free fetal DNA in maternal plasma. *Transfusion,* 53**,** 2892-8.

RIOS, CHAUDHURI, MALLINSON, SAUSAIS, GOMENSORO, G., HANNON, ROSENBERGER, POOLE, BURGESS, POGO & REID 2000. New genotypes in Fy(a− b−) individuals: nonsense mutations (Trp to stop) in the coding sequence of either FY A or FY B. *Br J Haematol,* 108**,** 448-454.

ROBINSON, J. T., THORVALDSDOTTIR, H., WINCKLER, W., GUTTMAN, M., LANDER, E. S., GETZ, G. & MESIROV, J. P. 2011. Integrative genomics viewer. *Nat Biotechnol,* 29**,** 24-6.

ROTHBERG, J. M., HINZ, W., REARICK, T. M., SCHULTZ, J., MILESKI, W., DAVEY, M., LEAMON, J. H., JOHNSON, K., MILGREW, M. J., EDWARDS, M., HOON, J., SIMONS, J. F., MARRAN, D., MYERS, J. W., DAVIDSON, J. F., BRANTING, A., NOBILE, J. R., PUC, B. P., LIGHT, D., CLARK, T. A., HUBER, M., BRANCIFORTE, J. T., STONER, I. B., CAWLEY, S. E., LYONS, M., FU, Y., HOMER, N., SEDOVA, M., MIAO, X., REED, B., SABINA, J., FEIERSTEIN, E., SCHORN, M., ALANJARY, M., DIMALANTA, E., DRESSMAN, D., KASINSKAS, R., SOKOLSKY, T., FIDANZA, J. A., NAMSARAEV, E., MCKERNAN, K. J., WILLIAMS, A., ROTH, G. T. & BUSTILLO, J. 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature,* 475**,** 348-52.

ROUBINET, F., DESPIAU, S., CALAFELL, F., JIN, F., BERTRANPETIT, J., SAITOU, N. & BLANCHER, A. 2004. Evolution of the O alleles of the human ABO blood group gene. *Transfusion,* 44**,** 707-15.

ROUBINET, F., KERMARREC, N., DESPIAU, S., APOIL, P. A., DUGOUJON, J. M. & BLANCHER, A. 2001. Molecular polymorphism of O alleles in five populations of different ethnic origins. *Immunogenetics,* 53**,** 95-104.

SACHIDANANDAM, R., WEISSMAN, D., SCHMIDT, S. C., KAKOL, J. M., STEIN, L. D., MARTH, G., SHERRY, S., MULLIKIN, J. C., MORTIMORE, B. J., WILLEY, D. L., HUNT, S. E., COLE, C. G., COGGILL, P. C., RICE, C. M., NING, Z., ROGERS, J., BENTLEY, D. R., KWOK, P. Y., MARDIS, E. R., YEH, R. T., SCHULTZ, B., COOK, L., DAVENPORT, R., DANTE, M., FULTON, L., HILLIER, L., WATERSTON, R. H., MCPHERSON, J. D., GILMAN, B., SCHAFFNER, S., VAN ETTEN, W. J., REICH, D., HIGGINS, J., DALY, M. J., BLUMENSTIEL, B., BALDWIN, J., STANGE-THOMANN, N., ZODY, M. C., LINTON, L., LANDER, E. S. & ALTSHULER, D. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature,* 409**,** 928-33.

SANGER, F., NICKLEN, S. & COULSON, A. R. 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A,* 74**,** 5463-7.

SANO, R., NAKAJIMA, T., TAKAHASHI, K., KUBO, R., KOMINATO, Y., TSUKADA, J., TAKESHITA, H., YASUDA, T., ITO, K., MARUHASHI, T., YOKOHAMA, A., ISA, K., OGASAWARA, K. & UCHIKAWA, M. 2012. Expression of ABO blood-group genes is dependent upon an erythroid cell-specific regulatory element that is deleted in persons with the B(m) phenotype. *Blood,* 119**,** 5301-10.

SCHMID, P., RAVENELL, K. R., SHELDON, S. L. & FLEGEL, W. A. 2012. DARC alleles and Duffy phenotypes in African Americans. *Transfusion,* 52**,** 1260-7.

SCHONEWILLE, H., VAN DE WATERING, L. M., LOOMANS, D. S. & BRAND, A. 2006. Red blood cell alloantibodies after transfusion: factors influencing incidence and specificity. *Transfusion,* 46**,** 250-6.

SCHWARZ, H. P. & DORNER, F. 2003. Karl Landsteiner and his major contributions to haematology. *Br J Haematol,* 121**,** 556-65.

SEATTLESEQANNOTATION137. 2016. Available: http://snp.gs.washington.edu/SeattleSeqAnnotation137 [Accessed 2016].

SELTSAM, A., WAGNER, F. F., GRUGER, D., GUPTA, C. D., BADE-DOEDING, C. & BLASCZYK, R. 2007. Weak blood group B phenotypes may be caused by variations in the CCAAT-binding factor/NF-Y enhancer region of the ABO gene. *Transfusion,* 47**,** 2330-5.

SHANKER, A. 2012. Genome research in the cloud. *OMICS,* 16**,** 422-8.

SHENDURE, J. & JI, H. 2008. Next-generation DNA sequencing. *Nat Biotechnol,* 26**,** 1135-45.

SIEBERT, P. D. & FUKUDA, M. 1986. Isolation and characterization of human glycophorin A cDNA clones by a synthetic oligonucleotide approach: nucleotide sequence and mRNA structure. *Proc Natl Acad Sci U S A,* 83**,** 1665-9.

SIMS, D., SUDBERY, I., ILOTT, N. E., HEGER, A. & PONTING, C. P. 2014. Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet,* 15**,** 121-32.

SINGLETON, B. K., GREEN, C. A., AVENT, N. D., MARTIN, P. G., SMART, E., DAKA, A., NARTER-OLAGA, E. G., HAWTHORNE, L. M. & DANIELS, G. 2000. The presence of an RHD pseudogene containing a 37 base pair duplication and a nonsense mutation in Africans with the Rh D-negative blood group phenotype. *Blood,* 95**,** 12-18.

SPRINGER, G. F. & HORTON, R. E. 1969. Blood group isoantibody stimulation in man by feeding blood group-active bacteria. *J Clin Invest,* 48**,** 1280-91.

STABENTHEINER, S., DANZER, M., NIKLAS, N., ATZMULLER, S., PROLL, J., HACKL, C., POLIN, H., HOFER, K. & GABRIEL, C. 2011. Overcoming methodical limits of standard RHD genotyping by next-generation sequencing. *Vox Sang,* 100**,** 381-8.

STORRY, J. R., CASTILHO, L., DANIELS, G., FLEGEL, W. A., GARRATTY, G., DE HAAS, M., HYLAND, C., LOMAS-FRANCIS, C., MOULDS, J. M., NOGUES, N., OLSSON, M. L., POOLE, J., REID, M. E., ROUGER, P., VAN DER SCHOOT, E., SCOTT, M., TANI, Y., YU, L. C., WENDEL, S., WESTHOFF, C., YAHALOM, V. & ZELINSKI, T. 2014. International Society of Blood Transfusion Working Party on red cell immunogenetics and blood group terminology: Cancun report (2012). *Vox Sang,* 107**,** 90-6.

STORRY, J. R., CASTILHO, L., DANIELS, G., FLEGEL, W. A., GARRATTY, G., FRANCIS, C. L., MOULDS, J. M., MOULDS, J. J., OLSSON, M. L., POOLE, J., REID, M. E., ROUGER, P., VAN DER SCHOOT, E., SCOTT, M., SMART, E., TANI, Y., YU, L. C., WENDEL, S., WESTHOFF, C., YAHALOM, V. & ZELINSKI, T. 2011. International Society of Blood Transfusion Working Party on red cell immunogenetics and blood group terminology: Berlin report. *Vox Sang,* 101**,** 77-82.

STORRY, J. R., JOUD, M., CHRISTOPHERSEN, M. K., THURESSON, B.,

AKERSTROM, B., SOJKA, B. N., NILSSON, B. & OLSSON, M. L. 2013. Homozygosity for a null allele of SMIM1 defines the Vel-negative blood group phenotype. *Nat Genet,* 45**,** 537-41.

STORRY, J. R. & OLSSON, M. L. 2004. Genetic basis of blood group diversity. *Br J Haematol,* 126**,** 759-71.

STORRY, J. R. & OLSSON, M. L. 2009. The ABO blood group system revisited: a review and update. *Immunohematology,* 25**,** 48-59.

TAKI, T. & KIBAYASHI, K. 2014. A simple ABO genotyping by PCR using sequence-specific primers with mismatched nucleotides. *Leg Med (Tokyo),* 16**,** 168-72.

TESTER, D. J., CRONK, L. B., CARR, J. L., SCHULZ, V., SALISBURY, B. A., JUDSON, R. S. & ACKERMAN, M. J. 2006. Allelic dropout in long QT syndrome genetic testing: A possible mechanism underlying false-negative results. *Heart Rhythm,* 3**,** 815-821.

THERMOFISHER. 2014. *The Ion Proton™ System* [Online]. Available: https://tools.thermofisher.com/content/sfs/brochures/CO06326_Proton_Spec_Sheet_FHR.pdf [Accessed 23 August 2016].

THERMOFISHER. 2015. *The Ion PGM System* [Online]. Thermo Fisher Scientific. Available: https://tools.thermofisher.com/content/sfs/brochures/PGM-Specification-Sheet.pdf [Accessed 23 August 2016].

THERMOFISHER. 2016. *Ion Torrent™ Next-Generation Sequencing Technology* [Online]. Available: https://www.thermofisher.com/uk/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-technology.html# [Accessed 23 August 2016].

THURESSON, B., HOSSEINI-MAAF, B., HULT, A. K., HUSTINX, H., ALAN CHESTER, M. & OLSSON, M. L. 2012. A novel B(weak) hybrid allele lacks three enhancer repeats but generates normal ABO transcript levels. *Vox Sang,* 102**,** 55-64.

TILLEY, L. & GRIMSLEY, S. 2014. Is Next Generation Sequencing the future of blood group testing? *Transfus Apher Sci,* 50**,** 183-8.

TOURNAMILLE, C., COLIN, Y., CARTRON, J. P. & LE VAN KIM, C. 1995. Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. *Nat Genet,* 10**,** 224-228.

TOURNAMILLE, C., KIM, C., GANE, P., CARTRON, J.-P. & COLIN, Y. 1995b. Molecular basis and PCR-DNA typing of the Fya/fyb blood group

polymorphism. *Human Genetics,* 95**,** 407-410.

TOURNAMILLE, C., LE VAN KIM, C., GANE, P., LE PENNEC, P. Y., ROUBINET, F., BABINET, J., CARTRON, J. P. & COLIN, Y. 1998. Arg89Cys Substitution Results in Very Low Membrane Expression of the Duffy Antigen/Receptor for Chemokines in Fyx Individuals. *Blood,* 92**,** 2147-2156.

URBANIAK, S. J. & GREISS, M. A. 2000. RhD haemolytic disease of the fetus and the newborn. *Blood Rev,* 14**,** 44-61.

VELDHUISEN, B., VAN DER SCHOOT, C. E. & DE HAAS, M. 2009. Blood group genotyping: from patient to high-throughput donor screening. *Vox Sang,* 97**,** 198-206.

VOELKERDING, K. V., DAMES, S. A. & DURTSCHI, J. D. 2009. Next-generation sequencing: from basic research to diagnostics. *Clin Chem,* 55**,** 641-58.

WAGNER, F. F. & FLEGEL, W. A. 2000. RHD gene deletion occurred in the Rhesus box. *Blood,* 95**,** 3662-3668.

WANG, J., WANG, W., LI, R., LI, Y., TIAN, G., GOODMAN, L., FAN, W., ZHANG, J., LI, J., ZHANG, J., GUO, Y., FENG, B., LI, H., LU, Y., FANG, X., LIANG, H., DU, Z., LI, D., ZHAO, Y., HU, Y., YANG, Z., ZHENG, H., HELLMANN, I., INOUYE, M., POOL, J., YI, X., ZHAO, J., DUAN, J., ZHOU, Y., QIN, J., MA, L., LI, G., YANG, Z., ZHANG, G., YANG, B., YU, C., LIANG, F., LI, W., LI, S., LI, D., NI, P., RUAN, J., LI, Q., ZHU, H., LIU, D., LU, Z., LI, N., GUO, G., ZHANG, J., YE, J., FANG, L., HAO, Q., CHEN, Q., LIANG, Y., SU, Y., SAN, A., PING, C., YANG, S., CHEN, F., LI, L., ZHOU, K., ZHENG, H., REN, Y., YANG, L., GAO, Y., YANG, G., LI, Z., FENG, X., KRISTIANSEN, K., WONG, G. K., NIELSEN, R., DURBIN, R., BOLUND, L., ZHANG, X., LI, S., YANG, H. & WANG, J. 2008. The diploid genome sequence of an Asian individual. *Nature,* 456**,** 60-5.

WAŚNIOWSKA, K., BLANCHARD, D., JANVIER, D., WANG, Z.-X., PEIPER, S. C., HADLEY, T. J. & LISOWSKA, E. 1996. Identification of the Fy6 epitope recognized by two monoclonal antibodies in the N-terminal extracellular portion of the Duffy antigen receptor for chemokines. *Molecular Immunology,* 33**,** 917-923.

WASNIOWSKA, K., LISOWSKA, E., HALVERSON, G. R., CHAUDHURI, A. & REID, M. E. 2004. The Fya, Fy6 and Fy3 epitopes of the Duffy blood group system recognized by new monoclonal antibodies: identification of a linear Fy3 epitope. *Br J Haematol,* 124**,** 118-122.

WATKINS, W. M. 2001. The ABO blood group system: historical background. *Transfusion Medicine,* 11**,** 243-265.

WESTER, E. S., STORRY, J. R. & OLSSON, M. L. 2011. Characterization of Jk(a+weak): a new blood group phenotype associated with an altered JK*01 allele. *Transfusion,* 51**,** 380-392.

WESTHOFF, C. M. & REID, M. E. 2004. Review: the Kell, Duffy, and Kidd blood group systems. *Immunohematology,* 20**,** 37-49.

WESTHOFF, C. M., VEGE, S., LOMAS-FRANCIS, C., HUE-ROYE K & PATEL, P. W. 2014. Identification of two new alleles, FY*B C.895G N A and FY*B C.179_180delCT, in the FY system associated with silencing of antigen expression. *Vox Sang,* 107(S1):195.

WETTERSTRAND, K. 2016. *DNA sequencing costs: Data from the NHGRI Genome Sequencing Programme* [Online]. National Human Genome Research Institute. Available: www.genome.gov [Accessed 21 Aug 2016].

WHEELER, D. A., SRINIVASAN, M., EGHOLM, M., SHEN, Y., CHEN, L., MCGUIRE, A., HE, W., CHEN, Y.-J., MAKHIJANI, V., ROTH, G. T., GOMES, X., TARTARO, K., NIAZI, F., TURCOTTE, C. L., IRZYK, G. P., LUPSKI, J. R., CHINAULT, C., SONG, X.-Z., LIU, Y., YUAN, Y., NAZARETH, L., QIN, X., MUZNY, D. M., MARGULIES, M., WEINSTOCK, G. M., GIBBS, R. A. & ROTHBERG, J. M. 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature,* 452**,** 872-876.

WILKINSON, K., HARRIS, S., GAUR, P., HAILE, A., ARMOUR, R., TERAMURA, G. & DELANEY, M. 2012. Molecular blood typing augments serologic testing and allows for enhanced matching of red blood cells for transfusion in patients with sickle cell disease. *Transfusion,* 52**,** 381-8.

WONG, L.-J. C. 2013. *Next Generation Sequencing: Translation to Clinical Diagnostics*.

YAMAMOTO, F. 2004. Review: ABO blood group system--ABH oligosaccharide antigens, anti-A and anti-B, A and B glycosyltransferases, and ABO genes. *Immunohematology / American Red Cross,* 20**,** 3-22.

YAMAMOTO, F., CLAUSEN, H., WHITE, T., MARKEN, J. & HAKOMORI, S. 1990a. Molecular genetic basis of the histo-blood group ABO system. *Nature,* 345**,** 229-33.

YAMAMOTO, F., MARKEN, J., TSUJI, T., WHITE, T., CLAUSEN, H. & HAKOMORI, S. 1990b. Cloning and characterization of DNA complementary

to human UDP-GalNAc: Fuc alpha 1----2Gal alpha 1----3GalNAc transferase (histo-blood group A transferase) mRNA. *J Biol Chem,* 265**,** 1146-51.

YAMAMOTO, F., MCNEILL, P. D. & HAKOMORI, S. 1992. Human histo-blood group A2 transferase coded by A2 allele, one of the A subtypes, is characterized by a single base deletion in the coding sequence, which results in an additional domain at the carboxyl terminal. *Biochem Biophys Res Commun,* 187**,** 366-74.

YAMAMOTO, F., MCNEILL, P. D. & HAKOMORI, S. 1995. Genomic organization of human histo-blood group ABO genes. *Glycobiology,* 5**,** 51-8.

YAMAMOTO, F., MCNEILL, P. D., YAMAMOTO, M., HAKOMORI, S., HARRIS, T., JUDD, W. J. & DAVENPORT, R. D. 1993. Molecular genetic analysis of the ABO blood group system: 1. Weak subgroups: A3 and B3 alleles. *Vox Sang,* 64**,** 116-9.

YAZDANBAKHSH, K., RIOS, M., STORRY, J. R., KOSOWER, N., PARASOL, N., CHAUDHURI, A. & REID, M. E. 2000. Molecular mechanisms that lead to reduced expression of Duffy antigens. *Transfusion,* 40**,** 310-320.

YUNIS, E. J., SVARDAL, J. M. & BRIDGES, R. A. 1969. Genetics of the Bombay phenotype. *Blood,* 33**,** 124-32.

ZHANG, X. H., HELLER, K. A., HEFTER, I., LESLIE, C. S. & CHASIN, L. A. 2003. Sequence information for the splicing of human pre-mRNA identified by support vector machine classification. *Genome Res,* 13**,** 2637-50.

ZIMMERMAN, P. A., WOOLLEY, I., MASINDE, G. L., MILLER, S. M., MCNAMARA, D. T., HAZLETT, F., MGONE, C. S., ALPERS, M. P., GENTON, B., BOATIN, B. A. & KAZURA, J. W. 1999. Emergence of FY*A(null) in a Plasmodium vivax-endemic region of Papua New Guinea. *Proc Natl Acad Sci U S A,* 96**,** 13973-7.

Appendices

Appendix A

Margaret Kenwright Young Scientist Award 2016 from the British Blood Transfusion Society (BBTS) for the following work:

Altayar, M.A., Madgett, T.E., Kiernan, M., Halawani, A.J., Avent, N.D. (2016). Next generation sequencing of *JK* (*SLC14A1*) gene reveals higher frequency of variant alleles, novel allele-defining SNPs (allele reference fingerprints) and reassignment of a purported *JKnull* allele. *Transfusion Medicine*, **26,** suppl. 2**,** 3-24.

Antigens of the Jk system are among causes of alloimmunisation. The number of alleles of this clinically significant blood group is increasing. Commercially available high-throughput Blood group genotyping (BGG) platforms are based on pre-defined polymorphisms. Next generation sequencing (NGS) circumvents this issue by operating in a discovery mode. We describe an NGS-based method for high-throughput, rapid and extensive genotyping of *JK*. Ion Torrent Personal Genome Machine (PGM^{TM}) was used to optimise the method for sequencing the entire *JK* gene plus flanking regions with Long-range PCR (LR-PCR). DNA libraries were prepared from 67 *JK* DNA samples with 3 LR-PCRs. Samples were loaded onto a 316 chip for sequencing. Sequencing data were aligned to the gene reference sequence and analysed by software packages, such as Ion Torrent Suite^{TM} plugins. Sanger sequencing of cDNA and cDNA clones was used for confirmation. The analysis of millions of reads generated with coverage depth of 700x showed: *JK*01W* allele (130G>A) is common (10/67) with normal antigenicity. Silencing polymorphism (810G>A) leading to purported *JK*B* null allele maintains AG splice site and Jk^{b} antigenicity (10/67). Intronic mutation analysis revealed novel allele-defining SNPs and a (G>Del) deletion, giving *JK*A*, *JK*B* and *JK*01W* novel reference allele sequence fingerprints, plus two new *JK*A* alleles were found. We found undescribed interactions between *JK*A-JK*B* allele sequences suggesting the *JK*01W* allele sequence arose from hybrid *JK*A-JK*B* gene sequence. C>T SNP, located 160bp upstream of the *JK*A/JK*B* critical polymorphism 838G>A, may lead to allelic dropout during BGG. Similarly, 53 *FY* and 45 *ABO* samples were processed revealing novel findings, such as 103G>A (Gly35Arg) in *ABO*.NGS can extensively genotype *JK* and other blood group genes including introns to define all existing and novel mutations. Thus, NGS will supplant array-based genotyping platforms and applying the discovery approach will refine the molecular basis of blood group genes.

N. D. Avent, T. E. Madgett, A. J. Halawani, **M. A. Altayar,** M. Kiernan, A. J. Reynolds, & X. Li. (2015). Next-generation sequencing: academic overkill or high-resolution routine blood group genotyping? *ISBT Science Series*, **10,** suppl. 1, 250-256.

Blood group genotyping (BGG) has been in routine clinical practice ever since the molecular determination of blood groups was achieved, during the early to mid 1990s. These early methods were dependent on allele-specific PCR to detect simple single nucleotide polymorphisms (SNPs) responsible for blood group expression. As our knowledge regarding the molecular background of blood groups advanced, so did the numbers of SNPs requiring detection which necessitated the switch from allele-specific PCR to array-based technology, notably slide and bead-based approaches. All these methods thus described are totally dependent on the predefined genetic basis of each SNP. Hybrid blood group genes (as found in RH, ABO and MNS systems) are difficult to define by all of these aforementioned techniques. With the arrival of cheap next-generation sequencing (NGS) approaches in the past 5 years, we have conducted long-range PCR (LRPCR) coupled with NGS determination of the major blood group genes. By analysis of these data using an Ion Torrent Personal Genome MachineTM (PGMTM), it is readily apparent that NGS can highly effectively be applied with high resolution to blood grouping at costs no more than current array-based platforms. However, the complexity of the data obtained need careful filtering for effective clinical utilization, but provides useful insight on the evolution of blood groups and their environmental impacts which will be of undoubted value as an academic exercise, but of minimal cost (if any) to the original testing.

M. Altayar, A. Halawani, M. Kiernan, A. Reynolds, N. Kaushik, T. Madgett & N. Avent (2013) Next Generation Sequencing of ABO, Duffy and Kidd Blood Group Genotyping. *Transfusion Medicine*, **23,** suppl. 2; 30-71.

Blood group Genotyping (BGG) has become well established in transfusion medicine. However, all current technologies are based on pre-existing knowledge of known polymorphisms. Next generation sequencing circumvents this requirement and adopts a discovery mode,which is important, as almost every new BGG project reveals new alleles. NGS has become high-throughput, rapid and accurate. Also, costs have significantly reduced in the past five years. In this pilot study, Ion Torrent Personal Genome Machine (PGM) sequencer was used to optimise and develop a reliable protocol for sequencing the entire Duffy, ABO and Kidd blood group genes including flanking regions. A DNA library was prepared from randomly selected DNA samples. Duffy, ABO and Kidd geneswere targeted by long-range PCR, enzymatic amplicon fragmentation before ligation with barcoded adapters and size selection. Templates were immobilised on beads, then clonally amplified using emulsion PCR. Sequencing revealed millions of reads that were then aligned to the reference gene sequences. Variants were visualised with two software packages, CLC workbench and Integrative Genomics Viewer (IGV). Initial Bioinformatics analysis of Duffy gene samples revealed all genotypes matched phenotype, however a number of SNPs (single nucleotide polymorphisms) were identified in exons 1,2 and intron1 and under investigation. We also identified that a significant number of the FY genes sequenced (5/12) encoded the previously described Ala100Thr mutation. The samples concerned

were FY (a-b+) (3 of 4) (1 homozygous and 2 heterozygous), FY (a+b-) (0 of 1) and FY (a+b+) (2/7) both heterozygous). We therefore conclude that the Ala100Thr mutation is much more frequent than previously described. We suggest that NGS will supplant other genotyping platforms in the near future and will potentially become the methodology of choice for genotyping patients and donors.

A. J. Halawani, **Altayar, M. A.,** M. Kiernan, N. Kaushik, A. Reynolds, T. Madgett & N. Avent (2013) Comprehensive Genotyping for Kell and Rh Blood Group Systems by Next-generation DNA Sequencing. *Transfusion Medicine* **23,** suppl. 2; 30-71.

Blood group genotyping (BGG) has emerged as a core technique in transfusion edicine and has impacted on the clinical management of multi-transfused patients. The vast ajority of these technical platforms require previous knowledge of the blood group polymorphisms under investigation. Next-generation sequencing (NGS) has emerged as a powerful replacement technology to genotyping single nucleotide polymorphisms (SNP), insertion/deletion (indels) and gene rearrangements. We have used NGS to define Rh and Kell alleles by amplification of the entire genes (KEL RHD,and RHCE). DNA was extracted from 14 random blood donor samples. Two primer pairs were designed to amplify the entire KEL gene using long-range polymerase chain reaction (LR-PCR). The sequencing library was constructed by fragmenting the DNA, ligating into barcoded adaptors and size selected using SPRIselect magnetic beads. The sequencing template was then immobilised to sphere particles that clonally amplified using emulsion PCR, emulsion breaking and enrichment for positive sphere particles. Finally, the sequencing reaction was loaded into a chip and sequenced on the Ion Torrent Personal Genome Machine (PGM) that generating 2·6 million reads. Data was analysed using CLC Genomics workbench Version 6·0.4. Data from the serological testing was confirmed by NGS. Two samples were typed serologically as K antigen and four as Kpb and this was confirmed by NGS as an initial genotype as [K antigen, Thr193Met (578C>T), heterozygous SNP 53·4% and 46·3%, respectively] and Kpb (Arg281, 841C and 842G). Moreover, one sample was found to be Kpa Arg281Trp (841C>T). Genotyping will be performed for all the antigens of Kell blood group system that encode the following antigens: (K/k, Kpa/Kpb/Kpc, Jsa /Jsb, K11/K17, K12, K13, K14/K24, VLAN/VONG, K18, K19, Km, K22, TOU, RAZ, KALT, KTIM, KYO/KYOR, KUCI, KANT, KASH, KELP, KETI and KHUL). NGS using LR-PCR approach offers a powerful technique enabling users to investigate comprehensive screening of all the Kell and Rh antigens. Similar approaches are in progress to define Rh alleles and that will reveal their variants in respect of the hybrid genes.

Altayar, M. A., Halawani, A. J., Kiernan, M., Reynolds, A. J., Kaushik, N., Madgett, T. E. & Avent, N. D. (2014). Extensive Genotyping of Blood Groups Duffy, Kidd and ABO by Next-generation Sequencing. *Vox Sang*, **107,** suppl. 1**,** 57-248.

The determination of the blood groups present in an individual attains a great clinical importance for the purposes of blood transfusion and transplantation. Blood group genotyping (BGG) has become well established in transfusion medicine. However, all current technologies are based on pre-defined knowledge of known polymorphisms. The approach of Next Generation Sequencing (NGS) circumvents this requirement and adopts a discovery mode, which is important, as almost every new BGG project discovers new alleles. NGS is capable of producing high-throughput, rapid and accurate data that results in extensive and detailed genotyping. Also, costs have considerably reduced in the past years. In this pilot study, Ion Torrent Personal Genome Machine

(PGMTM) sequencer was used to optimise and develop a reliable protocol for sequencing the entire Duffy (DARC), Kidd (SLC14A1) and ABO blood group genes including flanking regions. First, a DNA library was prepared from 12 randomly selected DNA samples. DARC, SLC14A1 and ABO genes were targeted by long-range PCR, enzymatic amplicon fragmentation before ligation with barcoded adapters and size selection. Templates were immobilised on beads, then clonally amplified using emulsion PCR. Then, up to 20 samples from FY and JK samples were selected with certain serology for sequencing. The reason for this is to seek the possibility that those serologically typed as negative for a particular blood group are actually genotypically weak positive. Sequencing revealed millions of reads with great coverage depth that were then aligned to the reference gene sequences. Variants were analysed and visualised with software packages, Ion Torrent SuiteTM plugins, CLC Genomic Workbench, Integrative Genomics Viewer (IGV) and SeattleSeq Annotation 138 website. Initial bioinformatics analysis of samples for the DARC and SLC14A1 genes revealed various single nucleotide polymorphisms (SNPs) in exons encoding for amino acid changes, such as (Gly42Asp and Ala100Thr in DARC) and (Asp280Asn, Ala270Ala and Glu44Lys in SLC14A1). In the FYsamples sequenced, a significant number (5/12) encoded the Ala100Thr mutation. In addition, the Allele JK*01W.01 (associated with Jka+w) was found with a frequency of 8%.A great number of polymorphisms (SNPs and Indel) were found in introns in JK samples (ranging from 52 to122) and in FY samples (ranging from 2 to 4). An example of one of these SNPs is chromosomal position 43319274 in the SLC14A1 gene, close to the splice site region of exon 8. Interestingly, sets of intronic polymorphisms present differently among samples with same phenotype. Therefore, the intronic polymorphisms are possibly unique to individuals or families. The ABO sequencing is in progress. We suggest that NGS shows the capability of the comprehensive sequencing of blood group genes and will supplant other genotyping platforms in the near future, becoming the potential methodology of choice for genotyping patients and donors.

Halawani, A. J., **Altayar, M. A.,** Kiernan, M., Reynolds, A. J., Kaushik, N., Madgett, T. E. & Avent, N. D. (2014). Can Next-generation DNA Sequencing Solve the RH Complexity for Genotyping? *Vox Sang*, **107,** suppl. 1**,** 57-248.

The Rh blood group system is considered as the most polymorphic blood group system and contains many variants. These variants may cause severe complications to patients, due to alloimmunisation from mismatching products of blood transfusion. Next-generation sequencing (NGS) is an intensely powerful technique capable of sequencing huge regions of the human genome. Here we sequenced the entire RHD and RHCE genes to genotype for the RhD and RhCE antigens, in order to provide a safer transfusion practice. DNA was extracted from random blood donors with known serology. By using long-range PCR (LR-PCR), four and three primer pairs were designed (giving PCR products in the range from 8 to 24 Kb) for RHD and RHCE, respectively. The sequencing libraries were then made by fragmenting the amplicons and ligating to adaptors using Ion XpressTM Plus Fragment Library Kit. Size selection was performed by SPRIselect magnetic beads. After that, the sequencing template was immobilised to beads, which possess a complementary strand to the adaptors, for clonal

amplification using emulsion PCR. Finally, the sequencing template was loaded onto a 316 chip and sequenced on the Ion Torrent Personal Genome MachineTM Sequencer. Millions of reads were generated and the data were analysed with CLC Genomics Workbench (Version 6.5). Sanger sequencing will be utilised in order to validate any variants from the sequencing data. NGS using the LR-PCR approach offers a crucial method assisting users to genotype many samples in a single run for detecting the Rh variants in depth. This will be extremely worthwhile to genotype the difficult alleles of Rh, especially hybrid genes, and will pave the way for the discovery of novel alleles. NGS for blood group alleles may represent a viable alternative to array and bead based platforms, and it is no more expensive and technically challenging on a per sample basis.

Altayar, M. A., Halawani, A. J., Kiernan, M., Madgett, T. E. & Avent, N. D. (2015). Complete Gene Sequencing of ABO Blood Group by Next-generation Sequencing. *Vox Sang,* **109,** suppl. 1**,** 1-379.

The ABO blood group system is the most clinical significant in blood transfusion and transplantation medicine. Due to naturally occurring antibodies, mismatched transfusion of blood can cause rapid transfusion reactions. ABO is one of the most complex and polymorphic blood group genes, with an ever-increasing number of variant alleles. These variant alleles not only affect the specificity of the enzymes but also the activity of the enzymes, which might result in a weak phenotype. Therefore the determination of ABO alleles is important for the safety of blood transfusion and transplantation medicine. Although high-throughput platforms have revolution ised the approach towards blood group genotyping (BGG), they are based on pre-defined polymorphisms, which are not suitable for the discovery of new alleles. Next Generation Sequencing (NGS) circumvents this requirement and operates in discovery mode, which is critical for emerging alleles. NGS is capable of producing comprehensive, high-throughput, rapid and accurate data resulting in extensive genotyping. ABO genotyping has frequently only focused on exons 6 and 7, neglecting the rest of the gene. Following our successful NGS-genotyping of blood group genes Duffy (DARC) and Kidd (SLC14A1), here we have used the Ion Torrent Personal Genome MachineTM (PGMTM) sequencer to optimise and develop a reliable protocol for sequencing the entire ABO blood group gene including flanking regions. In this pilot study, four long-range polymerase chain reactions (LR-PCR) were used to target the entire ABO gene plus over 5 kb upstream, regulatory regions (Promoter and CBF/NF-Y),anddownstream in 16 randomly selected genomic DNA samples. DNA libraries were prepared by enzymatic fragmentation, ligation of barcoded adapters and size selection, before clonal amplification of templates was achieved using emulsion PCR and then the samples were loaded onto a 316 chip for sequencing. Millions of reads with great coverage depth (100–1800x) were generated, which were then aligned to the reference gene sequence (NG_006669.1). These data were analysed and visualised with multiple software packages, such as CLC Genomic Workbench version 6.5. The serological phenotype data matched that of ABO genotyping. Bioinformatics analysis revealed a number of polymorphisms including single nucleotide polymorphisms (SNPs) and insertion/deletions (indels) distributed throughout exons, introns and the regulatory regions at around 4 kb upstream. Examples of amino acid changes due to the SNPs found in exons are those causing the differences

between *A* and *B* alleles, previously described in exon 7 (Arg176Gly, Gly235-Ser, Leu266Met and Gly268Ala), whilst other SNPs found in exons 3, 4 and 5 have been found to be of higher frequency in our samples than previously reported, including Arg63His (13/16 samples) and Ser74Pro (14/16 samples) and found in all ABO phenotypes. In addition, in two samples (of A and O phenotype), we showed the Trp181stop mutation, previously described only for the rare ABO*O.06 (O6) allele.
We suggest that NGS can provide a reliable approach to genotype ABO due to its powerful capabilities of comprehensive analysis and revealing novel alleles. NGS will supplant other genotyping platforms in the near future, becoming the potential methodology of choice for genotyping patients and donors for safe transfusion/ transplantation practice.

Halawani, A. J., **Altayar, M. A.,** Kiernan, M., Li, X., Madgett, T. E. & Avent, N. D. (2015). High Resolution Genotyping of the Rh Blood Group System by Next-generation Sequencing *Vox Sang*, **109,** suppl. 1**,** 1-379.

The RH blood group system is the most complicated blood group system due to encoding by two highly homologous genes, RHD and RHCE. Typing the antigens of this system by conventional serology is not very appropriate to distinguish between D positive, weak D and partial D. It is only possible to assign weak D and partial D alleles accurately using blood group genotyping (BGG). Here, 10 samples of the RH system were genotyped, 5 D positive and 5 weak D samples using a long-range PCR (LR-PCR) approach coupled with next generation sequencing (NGS). For every sample, both genes, RHD and RHCE, were amplified by LRPCR, with three amplicons for RHD and four amplicons for RHCE. Then the PCR products were fragmented ligated to barcoded adaptors and sequenced using NGS on an Ion Torrent PGMTM platform. We showed that LR-PCR for RHD and RHCE completely correlated with their corresponding genomic sequence. For the D positive samples, there were no obvious SNPs on the RHD exons. The 5 weak D samples have been identified as following; two weak D Type 1 (exon 6 809T>G Val270Gly), two weak D Type 2 (exon 9 1154G>C Gly385Ala) and one weak partial D 4.1 with [exon 1 48G>C (Trp16Cys), exon 4 602C>G (Thr201Arg), exon 5 667T>G (Phe223Val), exon 6 819G>A (silent)]. The LR-PCR method has confirmed that a novel heterozygous SNP, 208 C>T (Arg70Trp) in exon 2 is derived from the RHCE gene, although it had previously been identified by a Human Erythrocyte Antigen and Human Platelet Antigen panel as belonging to the RHD gene. More samples are currently being sequenced. Our approach, we believe, will facilitate the comprehensive genotyping of the antigens of the RH system, especially those with hybrid genes or insertions/deletions. Our method is able to demonstrate novel alleles by direct sequence analysis, a major drawback of current array-based BGG platforms.

Appendix B

Figure 5.3 with the raw trace from the 2100 Bioanalyzer® instrument

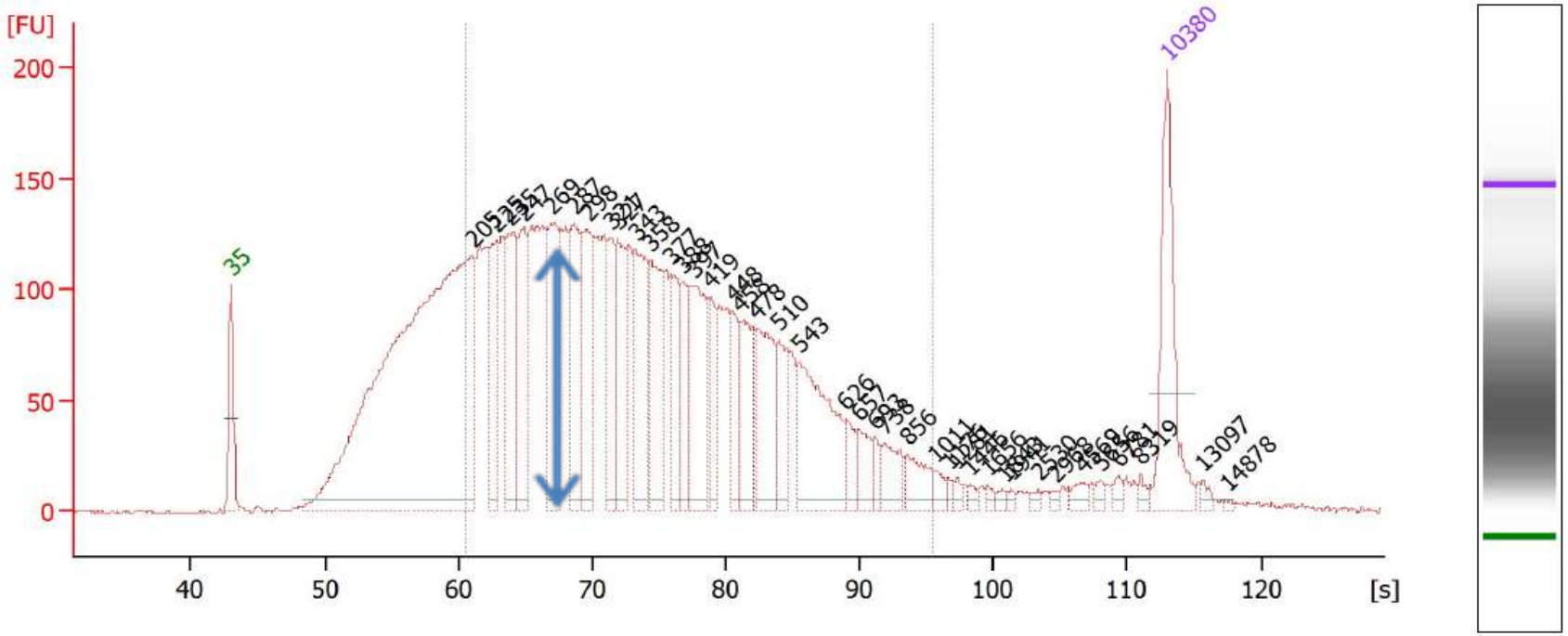


Figure 5.4 with the raw trace from the 2100 Bioanalyzer® instrument